Bayesian model selection for the glacial-interglacial cycle

Jake Carson†

Department of Mathematics, Imperial College London, UK

Michel Crucifix

Earth and Life Institute, Université catholique de Louvain, Belgium

Simon Preston

School of Mathematical Sciences, University of Nottingham, UK

Richard D. Wilkinson

School of Mathematics and Statistics, University of Sheffield, UK

Summary.

A prevailing viewpoint in paleoclimate science is that a single paleoclimate record contains insufficient information to discriminate between typical competing explanatory models. Here we show that by using SMC^2 (sequential Monte Carlo squared) combined with novel Brownian bridge type proposals for the state trajectories, it is possible to estimate Bayes factors to sufficient accuracy to be able to select between competing models, even with relatively short time series. The results show that Monte Carlo methodology and computer power have now advanced to the point where a full Bayesian analysis for a wide class of conceptual climate models is now possible. The results also highlight a problem with estimating the chronology of the climate record prior to further statistical analysis, a practice which is common in paleoclimate science. Using two datasets based on the same record but with different estimated chronologies, results in conflicting conclusions about the importance of the astronomical forcing on the glacial cycle, and about the internal dynamics generating the glacial cycle, even though the difference between the two estimated chronologies is consistent with dating uncertainty. This highlights a need for chronology estimation and other inferential questions to be addressed in a joint statistical procedure.

Keywords: Astronomical Forcing; Glacial Cycles; Model Comparison; Paleoclimate; Sequential Monte Carlo

†*Address for correspondence*: Jake Carson, Department of Mathematics, South Kensington Campus, Imperial College London, London, SW7 2AZ, UK.

Email: J.Carson@imperial.ac.uk

1. Introduction

Throughout the Pleistocene the Earth's climate has fluctuated between cold periods, in which glaciers expanded, and warm periods in which the glaciers retreated (Shackleton et al., 1984). The prevailing theory is that these glacial–interglacial (GIG) cycles are driven, or at least partly controlled, by variations in the *obliquity*, *precession*, and *eccentricity* of the Earth's orbit around the Sun, which affect the distribution of incoming solar radiation (or "insolation"). Obliquity, ε, is the angle between the equator and the orbital plane, and hence affects the distribution of insolation across seasons and latitudes. Precession is quantified by the angle about the Sun, ϖ, made by the point of perihelion (the point of the orbit when the Earth is closest to the Sun) and the vernal point (which marks the Spring equinox). At perihelion the Earth is anomalously close to the Sun when compared to a circular orbit, and so variation in precession affects the seasonal distribution of insolation. Eccentricity, e, measures how much the Earth's orbit deviates from being circular, and so modulates the effect of precession.

Milankovitch theory (Milankovitch, 1941) suggests that a positive anomaly of summer insolation in the Northern Hemisphere prevents the growth of ice sheets, a view supported by experiments with numerical simulations of the climate system (see Abe-Ouchi et al., 2013, for example). Such positive anomalies occur when perihelion is near the month of June, and/or when obliquity is anomalously large. As it is the combination of eccentricity, precession and obliquity that lead to high summer insolation, there is ambiguity about their respective roles (Crucifix, 2011). Consequently, many studies have examined the period and magnitude of each in order to see if any one can be considered to be the primary driver of the glacial-interglacial cycle. Opinion is varied and contradictory, with, for example, Huybers and Wunsch (2005) arguing for obliquity, Lisiecki (2010) for eccentricity, and Huybers (2011) for a combination of precession and obliquity. These studies each analysed features of the insolation signal using significance tests to assess whether the phases of each orbital component are significantly correlated with estimates of the glacial "termination times" (marking where individual glacial cycles finish), or whether the insolation signal was anomalously large at termination times. Differences in the details about how these tests were constructed appear to substantially affect the conclusions, with different studies finding different orbital characteristics being of primary importance.

It has also been recognised that despite the orbital control, the glacial-interglacial cycle is not entirely predictable (Imbrie and Imbrie, 1980; Raymo, 1997). This lack of predictability may emerge from interactions between the astronomical forcing and internal dynamics, in part because the quasi-periodic nature of the astronomical forcing hinders the robustness of the astronomical control (Crucifix, 2013; Mitsui and Aihara, 2014). An approach to study these effects is to model Earth's climate as a dynamical system forced by the variation in the astronomical parameters. To some extent the mechanical causes of the glacial-interglacial cycle have been investigated with numerical simulations of the physics of the atmosphere, ocean, ice-sheet (Abe-Ouchi et al., 2013), and carbon cycle (Brovkin

et al., 2012). The computational cost and number of parameters involved in such studies are often considerable, and in practice only a handful of simulations over the most recent cycles are performed. For these reasons, efforts to study changes in the regime and stability of the glacial-interglacial cycle have focused more on conceptual and phenomenological models, which typically involve just a few differential equations representing hypothesised relationships between different parts of the climate system. Atmospheric variability can be represented by stochastic processes, resulting in stochastic differential equation (SDE) models. Differences in the model structure and model parameters may yield shifts in the timing of the sequence of ice ages, especially if stochastic fluctuations due to atmosphere and ocean dynamics are properly accounted for (Mitsui and Crucifix, 2016), and so the challenge is to identify both what the appropriate forcing of the system is, and which is the best mathematical representation of the climate's internal dynamics.

The task of choosing between models is complicated by the nature of the data available. There are no reliable direct measurements of either the Earth's climate or of the extent of the glaciers before the 19th century. Climate proxy records, such as the ratio of oxygen isotopes ^{18}O and ^{16}O (referred to as δ^{18}O) measured in the calcite shells of foraminifera preserved in temporally stratified layers on the ocean floor, are instead used to construct estimates of climate and ice extent (Emiliani, 1955; Shackleton, 1967). These data are noisy, and contain uncertainties on both the measured climate proxy, and on the date relating to that measurement. Given this noise level, models representing very different properties or bifurcation structures may all be found to fit the data reasonably well when judged by eye (see Crucifix, 2012, for a recent account). As a consequence, a common viewpoint is that the information contained in a single proxy record is not sufficient to distinguish between the numerous proposed models (see, e.g., Roe and Allen, 1999).

In this paper we develop a fully Bayesian approach that simultaneously estimates model parameters, the relative contribution of each aspect of the orbital forcing, and chooses between models by estimating Bayes factors. The statistical difficulty in making inferences from partially and noisily observed trajectories of forced non-linear SDEs lies in the computation of various posterior quantities. Inference for SDEs is particularly challenging because the transition density, and therefore the likelihood function, is intractable (meaning it is not available in closed form). A powerful tool for time-structured problems with intractable likelihoods is the particle filter, and in this paper we employ the SMC^2 (sequential Monte Carlo squared) approach recently introduced by Chopin et al. (2013). This is a pseudo-marginal algorithm that embeds a particle filter within a sequential Monte Carlo algorithm to do joint state and parameter estimation. A major advantage of SMC^2 over competing methods, such as particle MCMC (Andrieu et al., 2010), is that it allows for easy estimation of the model evidence, which we exploit to provide estimates of the Bayes factors. A naive implementation of SMC^2 fails due to extreme particle degeneracy, but we show that by utilizing guided Brownian bridge type proposals more evenly distributed weights can be maintained. This is computationally

intensive even for phenomenological models, with the results in Section 4 each requiring 3-4 days on a single core. However, SMC is well suited to run in parallel. For example, using a Tesla K20 GPU these results can be obtained in 3-4 hours. A surprising result is that even though the Bayes factors are sensitive to both Monte Carlo error and the choice of prior distributions, the Bayes factors are sufficiently large, even for short time series of data, that we can still distinguish between the competing models.

Previous authors have also attempted model selection experiments for the GIG cycle, but with various limitations compared to our approach. Roe and Allen (1999) compared deterministic models plus autoregressive process noise using an F-test and found no support for any one model over any other. Feng and Bailer-Jones (2015) used Bayesian model selection to select between competing forcing functions over the Pleistocene, concluding that obliquity influences the termination times over the entire Pleistocene, and that precession also has explanatory power following the mid-Pleistocene transition. Their approach requires a tractable likelihood function, which heavily restricts the class of models that can be compared, in particular, ruling out the use of SDE models. As in the previously mentioned hypothesis tests, they also begin by discarding most of the data and using a summary consisting of just the termination times (~ 12 over the past 1 Myr), which is necessary as the low-order deterministic models used do not fit well to the complete dataset. They also only sample parameter values from the prior, leading to poor numerical efficiency. Finally, Kwasniok (2013) compares conceptual models over the last glacial period using the Bayesian information criterion. The likelihood of each model is estimated using an unscented Kalman filter (UKF) (Wan et al., 2000). Whilst this approach focussed on a smaller time horizon than our application, it can be applied using the data and models in this paper. However, the Gaussian approximation used by the UKF, whilst working well for filtering, is unproven for parameter estimation and model selection, and the particle filter offers a more natural approach for non-linear dynamical systems.

The approach presented here makes full use of the data rather than just the termination times, characterises parametric uncertainty rather than using plug-in estimates, and quantifies the evidence in favour of each model through Bayes factor estimates. We will show that it is possible to jointly estimate the state trajectory (a three dimensional vector over 800 time points), model parameters (up to 16 in one of the models), and estimate marginal likelihoods (allowing calculation of Bayes factors) even using relatively short time series of data, and that there is enough information in the data to choose between candidate conceptual models, including assessing the importance of the various orbital characteristics to the glacial–interglacial cycle.

The paper is structured as follows. In Section 2, we describe the data used, and the models of the astronomical forcing, the Earth's climate dynamics, and the proxy observations. Section 3 contains a description of the Bayesian approach, a brief review of the particle-filter, and we introduce our approach to inference, describing in detail how to avoid particle degeneracy using Brownian bridge

type proposals. In Section 4 we present a simulation study to assess the performance of the algorithms on synthetic data, and an analysis of a real δ^{18}O dataset. In Section 5 we offer some thoughts on the practical implementation of the particle filter methods for such problems, discuss the scientific conclusions, and suggest some future directions for research.

2. Data and Models

Our approach to understanding the dynamical behaviour of the paleoclimate involves four components: data consisting of paleoclimate records; models of the climate; drivers of the climate (such as CO_2 emissions or, more pertinently for paleoclimate, the orbital forcing); and a statistical model relating the foregoing three components. In this paper we develop the statistical methodology necessary for combining these components, which we hope will allow paleoclimate scientists to study hypotheses in a statistically rigorous way. That is to say, given some data and a selection of models, we show how to fit these models, and to assess which model is best supported by the data. Scientific aspects of the approach can, and we hope will, be improved upon by using different datasets and richer models.

2.1. Data

The ratio between the oxygen isotopes ^{18}O and ^{16}O, known as δ^{18}O, reflects a combination of effects associated with changes in ocean temperature and sea-level (Emiliani, 1955; Shackleton, 1967). Broadly speaking, larger values of δ^{18}O indicate a colder climate with greater ice volume. Time series of measured δ^{18}O recorded in the calcite shells of foraminifera are used as a proxy record for temperature and ice-extent, and provide a picture of the Earth's recent glaciations, particularly when combined with other information. The data we use are measurements of δ^{18}O from different depths in sediment cores extracted as part of the Ocean Drilling Programme (ODP). In climatology, a set of such measurements is known as a "record", and an average over multiple records is known as a "stack" (Imbrie et al., 1984). The δ^{18}O in deeper parts of a core correspond to climate conditions further back in time. Beyond monotonicity, there is unfortunately no simple relationship between core depth and age. This is because the accumulation of sediment results from a combination of complicated physical processes, including sedimentation (which occurs at variable rates), erosion, and core compaction. A model for the relationship between depth and age is known as an "age model", and application if such a model to a record is called "dating". A common strategy in developing an age model is to align features of records to important events observed in the core, such as magnetic reversals, whose dates are accurately known from other sources (Shackleton et al., 1990). In Huybers (2007), for instance, "age-control points" are identified in the core (such as glacial terminations, magnetic reversals, etc), and then ages for all the measurements are inferred from these control points, while accounting for compression using an involved heuristic process. Another common approach is to align features of the

δ^{18}O time series to aspects of the astronomical forcing, a process known as astronomical tuning. For example, in Lisiecki and Raymo (2005), ice ages are aligned with the predictions of the Imbrie and Imbrie (1980) model (linear relaxation with different relaxation times for glaciation and deglaciation forced by summer solstice insolation at 60° N).

The result of fitting an age model is a dataset $\{\tau_m, Y_m\}_{m=1}^{M}$ in which Y_m denotes the measurement of δ^{18}O at time τ_m. We use the ODP677 record (Shackleton et al., 1990), shown in Figure 1. The foraminifera here are of the benthic form, living in the deep ocean and therefore thought to be better representative of continental ice volume variations (see, though Elderfield et al., 2012). ODP677 has been dated both as part of an orbitally tuned scheme (Lisiecki and Raymo, 2005), and a non-orbitally tuned scheme (Huybers, 2007), giving two different dated records. It is widely known that in dating estimates from age models, the $\{\tau_m\}$, are highly uncertain, with accuracy believed to be of the order of 10 kyr (Huybers, 2007; Lisiecki and Raymo, 2005). Since investigating age models further is beyond the scope of this paper, here we treat τ_m as a given, but we discuss this assumption in Section 5. We will refer to the data from the orbitally tuned scheme as ODP677-f, where the 'f' denotes *forced*, and from the non-orbitally tuned scheme as ODP677-u, where the 'u' denotes *unforced*. We focus on the last 780 kyr of this record (the last magnetic reversal occurred 780 kya, allowing us to date the starting point accurately), which contains 363 observations, and use it to highlight issues surrounding double counting of the astronomical forcing.

Figure 1 about here.

2.2. The astronomical forcing

The amount of insolation hitting the top of the atmosphere at any point on Earth is a function of the hour angle (time in the day), the latitude, and the true solar longitude (i.e., time in the year). It also depends on the obliquity, precession and eccentricity of the Earth's orbit around the Sun, which vary over much longer time scales. *Obliquity* refers to the angle between the equator and the orbital plane, and controls the seasonal contrast. *Precession* of the point of perihelion (the point of the orbit when the Earth is closest to the Sun) with respect to the vernal point marking the spring equinox is quantified by the angle ϖ made by the two points about the sun. It determines when in the seasonal cycle the Earth is closest to the Sun, and causes the positive/negative anomaly insolation patterns sequentially across the different months of the year, thus controlling the length of the seasons. *Eccentricity*, e, measures how much the Earth's orbit deviates from being circular and modulates the effect of precession. Paleoclimatologists often transform eccentricity and precession, and refer instead to the *climatic precession*, $e \sin \varpi$, which is proxy for the effect of precession on the summer insolation in the Northern Hemisphere. By complementing climatic precession with $e \cos \varpi$ (proxy for spring insolation, termed here *coprecession*), insolation may effectively be computed at any time in the year and for any latitude (Berger, 1978).

For phenomenological models, which are not typically spatially resolved, the practice is to use some subset summary of the seasonal and spatial distribution of insolation. For example, Milankovitch theory (Milankovitch 1941; translation in Milankovitch 1998) asserts that the growth and shrinkage of ice sheets is controlled by summer insolation, typically at a reference latitude of 60° N, a quantity Milankovitch termed the caloric summer insolation. Other common summaries include the daily-mean insolation at summer solstice at 60° N (Imbrie and Imbrie, 1980), or at different times in the year (Saltzman and Maasch, 1990). These summaries may all be approximated as a linear combination of astronomical quantities, giving a forcing function, denoted $F(t;\boldsymbol{\gamma})$, as follows:

$$F(t;\boldsymbol{\gamma}) = \gamma_P \Pi_P(t) + \gamma_C \Pi_C(t) + \gamma_E E(t), \tag{1}$$

where $\Pi_P(t)$, $\Pi_C(t)$, and $E(t)$, are the normalised climatic precession ($e\sin\varpi$), coprecession ($e\cos\varpi$), and obliquity respectively. The parameter $\boldsymbol{\gamma} = (\gamma_P, \gamma_C, \gamma_E)^\top$ controls the linear combination. An algorithm to compute these quantities with sufficient accuracy for the late Pleistocene is provided in Berger (1978). More accurate, time indexed data are provided by Laskar et al. (2004) but the gain in accuracy is not critical in this context.

The geometry of ice sheets and snow line suggest that a positive insolation anomaly may lead to a greater ice volume change, than a negative one (Ruddiman, 2006). To account for this, some authors truncate the astronomical forcing to down-weight negative anomalies. Here, we introduce the truncation operator

$$f(x) = \begin{cases} x + \sqrt{4a^2 + x^2} - 2a & \text{if } x \leq 0 \\ x & \text{otherwise,} \end{cases}$$

where $a \geq 0$ is a constant that controls the minimum value. This function is used is used in model PP12 defined below (Paillard, 1998; Parrenin and Paillard, 2012).

2.3. Phenomenological models of climate dynamics

We consider three models of the climate dynamics. They were each originally proposed as low-order ordinary differential equations, with state vector $\boldsymbol{X}(t) = \big(X_{(1)}(t), ..., X_{(d)}(t)\big)^\top$, where d is the dimension of the model, with the first component $X_{(1)}$ representing global ice volume. The other components represent quantities such as glaciation state, or CO_2 concentration. In order to account for model errors, we convert the models into stochastic differential equations by the addition of a Brownian motion $\boldsymbol{W}(t)$. Note that alternative stochastic drivers (e.g., jump diffusions) could be considered, but in the absence of evidence that a more complex driver is required, we use Brownian motion as it is the simplest choice. These models were chosen as each models the glacial–interglacial cycle using a qualitatively different dynamical mechanism, as explained further below. For notational convenience we drop the explicit dependence of $\boldsymbol{X}$ and $\boldsymbol{W}$ on t.

Model SM91: (Saltzman and Maasch, 1991)

SM91 models glacial–interglacial cycles as a system of three SDEs,

$$\begin{aligned} \mathrm{d}X_{(1)} &= -\left(X_{(1)} + X_{(2)} + vX_{(3)} + F(\gamma_P, \gamma_C, \gamma_E)\right)\mathrm{d}t + \sigma_1 \mathrm{d}W_{(1)} \\ \mathrm{d}X_{(2)} &= \left(rX_{(2)} - pX_{(3)} - sX_{(2)}^2 - X_{(2)}^3\right)\mathrm{d}t + \sigma_2 \mathrm{d}W_{(2)} \\ \mathrm{d}X_{(3)} &= -q\left(X_{(1)} + X_{(3)}\right)\mathrm{d}t + \sigma_3 \mathrm{d}W_{(3)} \end{aligned}$$

in which variables $X_{(2)}$ and $X_{(3)}$ represent CO2 concentration in the atmosphere and deep-sea ocean temperature, respectively. The model is an expression of the hypothesis that carbon-cycle effects are critical for the emergence of glacial cycles. Hence the non-linear terms, which are responsible for the oscillation, are present in the second equation only. It uses basic knowledge about the temperature sensitivity to CO_2 and the astronomical forcing (encoded in the equation for $\mathrm{d}X_{(1)}$), but uses heuristic relationships for the slow feedback loops between CO_2, ocean temperature and ice growth. Uncertainty about sensitivity to CO_2 is implicitly accounted for by the scaling relationship between the variable $X_{(2)}$ and actual CO_2, and the uncertainty about forcing effects is encoded in prior distributions for γ_P, γ_C and γ_E. Parameter v determines the uncertain relationship between ice growth and ocean circulation (which control ocean temperature) though we will assume it to be positive (warming oceans cause ice to melt). The parameters r, p and s in equation $\mathrm{d}X_{(1)}$ encode the hypothesis that ice ages cycles are caused by an internal instability of the ocean-carbon cycle system (Saltzman, 1988). Parameter q sets the response time scale of ocean temperature. SM91 is non-dimensional with a reference value of 10 kyr for t.

Model T06: (Tziperman et al., 2006)

T06 is an example of a "hybrid" model coupling $X_{(1)}$, which is governed by a stochastic differential equation, to a binary variable $X_{(2)}$ indicating whether sea ice extent exceeds some critical threshold.

$$\begin{aligned} \mathrm{d}X_{(1)} &= \left(\left(p_0 - KX_{(1)}\right)\left(1 - \alpha X_{(2)}\right) - \left(s + F(\gamma_P, \gamma_C, \gamma_E)\right)\right)\mathrm{d}t + \sigma_1 \mathrm{d}W_{(1)} \\ X_{(2)} &: \quad \text{switches from 0 to 1 when } X_{(1)} \text{ exceeds some threshold } T_u \\ X_{(2)} &: \quad \text{switches from 1 to 0 when } X_{(1)} \text{ decreases below some threshold } T_l \end{aligned}$$

When T_u and T_l are suitably chosen, the resulting dynamics are that of a relaxation oscillation: $X_{(1)}$ tends either to increase or decrease depending on the state of $X_{(2)}$, and the trend reverses as $X_{(1)}$ crosses a threshold causing $X_{(2)}$ to switch state. As in SM91, the oscillation is further controlled by the astronomical forcing, allowing for synchronisation effects (Tziperman et al., 2006). The original motivation for T06 was to suggest a critical role for Arctic sea-ice cover (Ashkenazy and Tziperman, 2004); a positive sea-ice cover anomaly reduces the amount of snow fall on Northern Hemisphere ice caps, thereby acting *negatively* on the growth of ice (hence the $-X_{(2)}$ term in the equation for $\mathrm{d}X_{(1)}$), and vice-versa. Parameter p_0 is the precipitation rate with no sea ice, K is a growth rate constant,

s is an ablation constant, and α is the relative area of sea ice. Ice volume is expressed in units of $10^{15}\mathrm{m}^3$, K in units of kyr^{-1}, and p_0 and s in units of $10^6\mathrm{m}^3\mathrm{s}^{-1}$.

Model PP12: (Parrenin and Paillard, 2012)

PP12 is also a hybrid model, with $X_{(2)}$ now representing a hidden climatic state, that may either be "glaciation" (0) or "deglaciation" (1).

$$\begin{aligned} \mathrm{d}X_{(1)} &= -(\gamma_P \Pi_P^\dagger + \gamma_C \Pi_C^\dagger + \gamma_E E - a_g + (a_g + a_d + X_{(1)}/\tau)X_{(2)})\mathrm{d}t + \sigma_1 \mathrm{d}W_{(1)}, \\ X_{(2)} &: \quad \text{switches from 0 to 1 when } F(\kappa_P, \kappa_C, \kappa_E) \text{ is less than some threshold } v_\mathrm{l} \\ X_{(2)} &: \quad \text{switches from 1 to 0 when } F(\kappa_P, \kappa_C, \kappa_E) + X_{(1)} \text{ is greater than some threshold } v_\mathrm{u} \end{aligned}$$

where $\Pi_P^\dagger$ and $\Pi_C^\dagger$ are transformed precession and coprecession components defined to be

$$\begin{aligned} \Pi_P^\dagger &= (f(\Pi_P) - 0.148)/0.808 \\ \Pi_C^\dagger &= (f(\Pi_C) - 0.148)/0.808. \end{aligned}$$

This discrete distinction of climatic states is given an empirical justification in the data analysis by Imbrie et al. (2011). During the glaciation phase, ice volume trends upwards and is linearly controlled by the truncated insolation. The deglaciation phase is simply a relaxation towards low ice volume. The model assumes that the switch from deglaciation to glaciation is controlled by insolation alone, whereas the glaciation-deglaciation switch is determined by a condition on glaciation and insolation. This contrasts with T06, where the state changes are determined only by the system state. Consequently, a constant astronomical forcing cannot induce spontaneous oscillations in PP12. Parameters a_g and a_d are growth and ablation constants respectively. Ice volume is expressed as sea-level equivalent in meters, γ_P, γ_C, γ_E, a_g, and a_d in units of $\mathrm{m(kyr)}^{-1}$, τ in kyr, and κ_P, κ_C, κ_E, v_l, and v_u in meters.

A comparison of the ice volume generated from the deterministic version of each model is shown in Figure 2, using the parameters suggested in the original publications. Each model captures the broad structure of the glacial-interglacial cycle. These plots are not precise reproductions of the figures in the original publications, due to differences in the initial conditions and astronomical solutions. Note that each model was tuned using a different dataset, and so, for example, SM91 has seven cycles in 780 kyr rather than eight.

Figure 2 about here.

2.4. Statistical observation model

The final modelling ingredient is a statistical model relating the state variables in the dynamical climate models, $\boldsymbol{X}(t)$, to the dataset. We assume that the data are of the form $\{\tau_m, Y_m\}_{m=1}^M$, where

τ_m is the estimated age and Y_m the measured proxy of the m^{th} data point/slice. We use the model

$$Y_m \sim \mathcal{N}(D + \boldsymbol{H}^\top \boldsymbol{X}_m, \sigma_y^2),$$

where we define $\boldsymbol{X}_m = \boldsymbol{X}(\tau_m)$. As with the choice of stochastic driver, we have assumed the simplest observation model, and more complex observation models could be considered. Here, we use $\boldsymbol{H} = (C, 0, \ldots, 0)^\top$, so that Y_m is a scaled and shifted version of the value $X_{(1)}(\tau_m)$, the ice volume in the underlying dynamical model. However, vector observations can be used at no additional cost or complication to the methodology, allowing observations of other proxies if desired.

2.5. Prior distributions

A Bayesian approach requires specification of prior distributions for the parameters in each of the models. Milankovitch theory suggests that a positive northern hemisphere insolation anomaly in summer increases ablation over the ice sheets, giving a negative contribution to the ice volume derivative (Berger and Loutre, 2004). This requires that γ_P and γ_E be positive, and so we use exponential prior distributions on these parameters to allow the system to be weakly forced. Whether having more insolation in spring at the expense of autumn should result in a positive or negative contribution to ice accumulation is undetermined, and so we use a zero-mean Gaussian prior distribution for γ_C, as this is symmetric about zero, which indicates Summer solstice insolation. Beyond these choices, specifying prior distributions on physical grounds is difficult, as many of the model parameters do not represent measurable quantities. In general we select moderately informative prior distributions that discourage non-oscillating regimes, excessively small or large periods, and numerical instabilities. The complete set of prior distributions for all three models is given in Table 1. Sensitivity to the choice of prior distributions is investigated in Section 4.

Since we are performing model comparison, we have made the prior distributions as consistent as possible across models. For instance, since we scale the output from each model, the observation error variance, σ_y^2, is comparable, and so we use the same prior distribution for σ_y^2 in each model. Care needs to be taken when selecting prior distributions for the astronomical forcing terms as each model has a different scale for ice volume. From trial runs the range of $X_{(1)}$ in PP12 is approximately 2.5 times that of T06, and 50 times that of SM91. Additionally, SM91 has a reference value for t of 10 kyr (as opposed to 1 kyr in T06 and PP12), suggesting a ratio of 1:2:5 for the astronomical forcing parameters between SM91, T06, and PP12 respectively. This relationship is represented in the prior distributions. Since this ratio is only approximate, we investigate sensitivity to this scaling rule in Section 4.

Table 1 about here.

3. Methodology

Our primary aim is model comparison, in particular to answer the question: given a collection of competing models $\{\mathcal{M}_l\}_{l=1}^L$, which is best supported by the data? The Bayes factor (BF) for comparing two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ say, is the ratio of their evidences

$$B_{12} = \frac{\pi(Y_{1:M} \mid \mathcal{M}_1)}{\pi(Y_{1:M} \mid \mathcal{M}_2)},$$

where $Y_{1:M} = (Y_1, \ldots, Y_M)$ and where $\pi(Y_{1:M} \mid \mathcal{M}_l)$ is the evidence for model $\mathcal{M}_l$ (Jeffreys, 1939; Kass and Raftery, 1995). The Bayes factor summarises the strength of evidence in the data in support of one model over another, and is the ratio of the posterior to the prior odds in favour of $\mathcal{M}_1$ over $\mathcal{M}_2$. If the prior probabilities for each model are equal, then the Bayes factor is the ratio of the posterior model probabilities. There are many other approaches to model comparison, many of which are based on measures of predictive accuracy such as cross-validation (see Vehtari and Ojanen (2012) and Gelman et al. (2014)). Most of these approaches cannot be used in problems with complex data dependencies, such as serial correlation. Ando and Tsay (2010) describe an approach that can be used, but which requires calculation of the Hessian matrix of the log-likelihood. Additionally, many of these methods require evaluation of quantities which may not be available for complex state space models. For these reasons, and because our focus is on model comparison rather than on evaluating predictive accuracy, we prefer to use Bayes factors to perform model comparison.

Secondary aims of our analysis include parameter estimation and filtering, which in this context are often called calibration and climate reconstruction. Calibration is the process of finding the posterior distribution of the model parameters $\pi(\boldsymbol{\theta}_l \mid Y_{1:M}, \mathcal{M}_l)$, where $\boldsymbol{\theta}_l$ is the parameter for model $\mathcal{M}_l$, and filtering is finding the distribution of the state variables $\pi(\boldsymbol{X}_{1:M} \mid Y_{1:M}, \boldsymbol{\theta}_l, \mathcal{M}_l)$. These three problems are of different levels of difficulty. Filtering is the most straightforward, but is not simple as for non-linear or non-Gaussian models, direct calculation of the filtering distributions is not possible, and so we must instead rely upon approximations. Calibration requires that we integrate out the dependence on the state variables $\boldsymbol{X}_{1:M}$,

$$\pi(\boldsymbol{\theta}_l \mid Y_{1:M}, \mathcal{M}_l) = \int \pi(\boldsymbol{\theta}_l, \boldsymbol{X}_{1:M} \mid Y_{1:M}, \mathcal{M}_l) \mathrm{d}\boldsymbol{X}_{1:M},$$

and hence, is considerably more difficult than filtering. Finally, model selection requires integrating out the dependence on $\boldsymbol{\theta}_l$,

$$\pi(Y_{1:M} \mid \mathcal{M}_l) = \int \pi(\boldsymbol{\theta}_l \mid \mathcal{M}_l) \int \pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}_l, \mathcal{M}_l\right) \pi(Y_{1:M} \mid \boldsymbol{\theta}_l, \boldsymbol{X}_{1:M}, \mathcal{M}_l) \mathrm{d}\boldsymbol{X}_{1:M} \mathrm{d}\boldsymbol{\theta}_l,$$

and is thus even more difficult than calibration.

The development of Monte Carlo methodology for solving these three problems for state space models reflects this hierarchy of difficulty. Particle filter methodology, first proposed in the 1990s (Gordon et al., 1993), is able to solve the general filtering problem adequately as long as the dimension

of $\boldsymbol{X}_{1:M}$ is not too large. Whereas the calibration problem has only begun to be satisfactorily answered more recently, with the development of pseudo-marginal methods such as particle-MCMC (Andrieu et al., 2010). Calculating the model evidence is, however, still very much an open problem.

Here, we demonstrate how the recently introduced SMC^2 algorithm (Chopin et al., 2013) can be used to estimate model evidences. The approach relies upon the following identities decomposing the evidence:

$$\pi(Y_{1:M}) = \pi(Y_1) \prod_{m=2}^{M} \pi(Y_m \mid Y_{1:m-1}), \tag{2}$$

and

$$\pi(Y_m \mid Y_{1:m-1}) = \int \pi(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid Y_{1:m-1}) \mathrm{d}\boldsymbol{\theta}, \tag{3}$$

where we have dropped the dependence on $\mathcal{M}_l$ from the notation. SMC^2 can be used to sample from $\pi(\boldsymbol{\theta} \mid Y_{1:m-1})$, find unbiased estimates of $\pi(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta})$, and, by plugging these estimates into Equations (2) and (3), obtain an estimate of the model evidence, from which we can estimate the Bayes factors.

3.1. Estimating Model Evidence Using SMC^2

Sequential Monte Carlo algorithms (SMC) (Del Moral et al., 2006) are population-based sampling methods that aim to sample from some target distribution, π_M, by sampling from a series of intermediary distributions, $\{\pi_m\}_{m=1}^{M}$, that are chosen to gradually 'close-in' on the target distribution. SMC uses a weighted collection of particles to approximate each distribution, and sequentially updates the weights and the particles in such a way that the normalising constant of each distribution can be estimated. A common choice for the sequence of distributions is to add a single data point at a time, so that the intermediary distributions are $\pi(\boldsymbol{\theta} \mid Y_{1:m})$ or $\pi(\boldsymbol{X}_{1:m} \mid Y_{1:m})$, for example.

One of the earliest SMC algorithms is the particle filter (PF) (Gordon et al., 1993), which samples from the sequence of filtering distributions $\pi_m(\boldsymbol{X}_{1:m}) = \pi(\boldsymbol{X}_{1:m} \mid Y_{1:m}, \boldsymbol{\theta})$, and is described in Algorithm 1. The basic idea is that at initialisation, a sample of N_x particles are sampled from some initial proposal density $r_1(\boldsymbol{X}_1 \mid Y_1, \boldsymbol{\theta})$, and given importance weight $\pi(\boldsymbol{X}_1, Y_1 \mid \boldsymbol{\theta}) \diagup r_1(\boldsymbol{X}_1 \mid Y_1, \boldsymbol{\theta})$. These particles are then repeatedly resampled using a multinomial scheme, propagated via some arbitrary proposal distribution $r_m(\boldsymbol{X}_m \mid \boldsymbol{X}_{1:m-1}, Y_{1:m}, \boldsymbol{\theta})$, and reweighted accordingly, so that for each successive iteration the particles are a weighted sample of the posterior $\pi(\boldsymbol{X}_{1:m} \mid Y_{1:m}, \boldsymbol{\theta})$. Details of the resampling and the proposal distributions, r_m, are discussed in Section 3.2, and further details can be found in Doucet and Johansen (2009).

An important aspect of the PF is that an unbiased estimate of the normalising constant $\pi(Y_{1:m} \mid \boldsymbol{\theta})$ can be estimated from the unnormalised weights in each iteration of the algorithm, using

$$\hat{\pi}(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta}) = \frac{1}{N_x} \sum_{k=1}^{N_x} \omega_m^{(k)}(\boldsymbol{X}_{1:m}^{(k)})$$

Algorithm 1 Particle filter targeting $\pi\left(\boldsymbol{X}_{1:M} \mid Y_{1:M}, \boldsymbol{\theta}\right)$.

for $k = 1, ..., N_X$ **do**

Sample $\boldsymbol{X}_1^{(k)} \sim r_1\left(\boldsymbol{X}_1 \mid Y_1, \boldsymbol{\theta}\right)$.

Set the importance weight

$$\omega_1^{(k)} = \frac{\pi\left(\boldsymbol{X}_1^{(k)} \mid \boldsymbol{\theta}\right) \pi\left(Y_1 \mid \boldsymbol{X}_1^{(k)}, \boldsymbol{\theta}\right)}{r_1\left(\boldsymbol{X}_1^{(k)} \mid Y_1, \boldsymbol{\theta}\right)}.$$

end for

Normalise the weights. For $k = 1, ..., N_X$

$$\Omega_1^{(k)} = \frac{\omega_1^{(k)}}{\sum_{i=1}^{N_X} \omega_1^{(i)}}.$$

for $m = 2, ..., M$ **do**

for $k = 1, ..., N_X$ **do**

Sample ancestor particle index $a_{m-1}^{(k)}$ according to weights $\Omega_{m-1}^{(1:N_X)}$.

Sample $\boldsymbol{X}_m^{(k)} \sim r_m\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, Y_m, \boldsymbol{\theta}\right)$.

Extend particle trajectory $\boldsymbol{X}_{1:m}^{(k)} = \left\{\boldsymbol{X}_{1:m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{X}_m^{(k)}\right\}$.

Set the importance weight

$$\omega_m^{(k)} = \frac{\pi\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{\theta}\right) \pi\left(Y_m \mid \boldsymbol{X}_m^{(k)}, \boldsymbol{\theta}\right)}{r_m\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, Y_m, \boldsymbol{\theta}\right)}. \tag{4}$$

end for

Normalise the weights. For $k = 1, ..., N_X$

$$\Omega_m^{(k)} = \frac{\omega_m^{(k)}}{\sum_{i=1}^{N_X} \omega_m^{(i)}}.$$

end for

as an approximation to Equation (3), and then plugging these estimates into Equation (2) (Del Moral, 2004)

$$\hat{\pi}(Y_{1:M} \mid \boldsymbol{\theta}) = \hat{\pi}(Y_1) \prod_{m=2}^{M} \hat{\pi}(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta}). \tag{5}$$

In Andrieu and Roberts (2009), it was shown that using these unbiased estimates of the likelihood in other Monte Carlo algorithms can lead to valid Monte Carlo algorithms (termed pseudo-marginal algorithms) for performing parameter estimation. For example, PMCMC (Andrieu et al., 2010) uses the PF within an MCMC algorithm, and SMC^2 (Chopin et al., 2013) uses a PF embedded within an SMC algorithm, both with the aim of finding $\pi(\boldsymbol{\theta} \mid Y_{1:M})$. We concentrate on the latter as it allows for estimation of BFs.

The SMC^2 algorithm (Chopin et al., 2013) embeds the particle filter within an SMC algorithm targeting the sequence of posteriors

$$\pi_0 = \pi(\boldsymbol{\theta}), \qquad \pi_m = \pi(\boldsymbol{\theta}, \boldsymbol{X}_{1:m} \mid Y_{1:m}),$$

for $m = 1, \ldots, M$. This is achieved by initially sampling N_θ parameter particles, $\{\boldsymbol{\theta}^{(n)}\}_{n=1}^{N_\theta}$, from the prior. To each $\boldsymbol{\theta}^{(n)}$, we attach a PF of N_x particles, i.e., at iteration m the PF $\{\boldsymbol{X}_{1:m}^{(k,n)}, \Omega_m^{(k,n)}\}_{k=1}^{N_x}$ is associated with $\boldsymbol{\theta}^{(n)}$, where $\Omega_m^{(k,n)}$ are the normalised weights in Algorithm 1. From this PF we can obtain an unbiased estimate of the marginal likelihood $\pi(Y_{1:m} \mid \boldsymbol{\theta}^{(n)})$ via Equation (5). To assimilate the next observation Y_{m+1}, we first extend the PF for the state particles to $\{\boldsymbol{X}_{1:m+1}^{(k,n)}, \Omega_{m+1}^{(k,n)}\}_{k=1}^{N_x}$, and then estimate $\pi(Y_{1:m+1} \mid \boldsymbol{\theta}^{(n)})$ and so on. Particle degeneracy occurs when the weighted particle approximation is dominated by just a few particles (i.e., a few have comparatively large weights), and is monitored by calculating the effective sample size (ESS)

$$\text{ESS} = \left(\sum_{i=1}^{N_\theta} \left(W_m^{(i)} \right)^2 \right)^{-1},$$

where $\left\{ W_m^{(i)} \right\}_{i=1}^{N_\theta}$ are the normalised weights in population m. When the ESS falls below some threshold (usually $N_\theta/2$) the particles are resampled to discard low-weight particles. However, resampling can lead to too few unique particles in the parameter space. Particle diversity is improved by running a PMCMC algorithm that leaves $\pi(\boldsymbol{\theta}, \boldsymbol{X}_{1:m} \mid Y_{1:m})$ invariant, specifically the particle marginal Metropolis-Hastings (PMMH) algorithm (Andrieu et al., 2010). The full details of the SMC^2 algorithm are presented in Algorithm 2, with theoretical justification in Chopin et al. (2013).

The model evidence $\pi(Y_{1:M})$ can be decomposed according to Equation (2), and in each iteration of the SMC^2 algorithm, the term

$$\hat{\pi}(Y_m \mid Y_{m-1}) = \sum_{n=1}^{N_\theta} W^{(n)} \hat{\pi}(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta}^{(n)})$$

provides an estimate of $\pi(Y_m \mid Y_{m-1})$. An estimate of the model evidence $\pi(Y_{1:M})$ is then obtained from (2) with $\hat{\pi}(Y_m \mid Y_{1:m-1})$ substituted for $\pi(Y_m \mid Y_{1:m-1})$.

Algorithm 2 SMC^2 algorithm targeting $\pi(\theta, X_{1:M} \mid Y_{1:M})$.

for $n = 1, ..., N_\theta$ **do**

Sample $\boldsymbol{\theta}^{(n)}$ from the prior distribution, $\pi(\boldsymbol{\theta})$.

Set the importance weight

$$W_0^{(n)} = \frac{1}{N_\theta}.$$

end for

for $m = 1, ..., M$ **do**

if ESS$< \frac{N_\theta}{2}$ **then**

for $n = 1, ..., N_\theta$ **do**

Sample $\boldsymbol{\theta}^{*(n)}$ and $\boldsymbol{X}_{1:m-1}^{*(1:N_X,n)}$ from $\boldsymbol{\theta}^{(1:N_\theta)}$ and $\boldsymbol{X}_{1:m-1}^{(1:N_X,1:N_\theta)}$, according to weights $W_{m-1}^{(1:N_\theta)}$.

Sample $\boldsymbol{\theta}^{**(n)}$ and $\boldsymbol{X}_{1:m-1}^{**(1:N_X,n)}$ from a PMMH algorithm targeting $\pi(\boldsymbol{\theta}, \boldsymbol{X}_{1:m-1} \mid Y_{1:m-1})$ initialised with $\boldsymbol{\theta}^{*(n)}$ and $\boldsymbol{X}_{1:m-1}^{*(1:N_X,n)}$.

end for

Set $\boldsymbol{\theta}^{(1:N_\theta)} = \boldsymbol{\theta}^{**(1:N_\theta)}$ and $\boldsymbol{X}_{1:m-1}^{(1:N_X,1:N_\theta)} = \boldsymbol{X}_{1:m-1}^{**(1:N_X,1:N_\theta)}$.

Set the importance weights

$$W_{m-1}^{(n)} = \frac{1}{N_\theta} \text{ for } n = 1, ..., n_\theta.$$

end if

for $n = 1, ..., N_\theta$ **do**

Sample $\boldsymbol{X}_{1:m}^{(1:N_X,n)}$ by performing iteration m of the particle filter, and record estimates of $\hat{\pi}\left(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta}^{(n)}\right)$ and $\hat{\pi}\left(Y_{1:m} \mid \boldsymbol{\theta}^{(n)}\right)$.

Set the importance weights

$$w_m^{(n)} = W_{m-1}^{(n)} \hat{\pi}\left(Y_m \mid Y_{1:m-1}, \boldsymbol{\theta}^{(n)}\right).$$

end for

Evaluate

$$\hat{\pi}(Y_m \mid Y_{1:m-1}) = \sum_{i=1}^{N_\theta} w_m^{(i)}$$

Normalise the weights

$$W_m^{(n)} = \frac{w_m^{(n)}}{\sum_{i=1}^{N_\theta} w_m^{(i)}} \quad \text{for} \quad n = 1, ..., N_\theta.$$

end for

3.2. Guided proposals

A further difficulty arises as the transition densities $\pi(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta})$ are not available in closed form for the models of interest, suggesting that we need to choose the particle proposal distributions, $\{r_m\}$, so that the transition density cancels from the importance weights. This can be achieved by setting $r_m = \pi(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta})$ in Equation (4), so that proposals are just simulations from the model. However, this choice will typically lead to particle degeneracy if too many of the proposals end up being far from the observations, due to the light Gaussian tails in the observation model. Resampling the state particles ensures that important particles are propagated forward, which can improve the approximation in later iterations. Multinomial resampling is the most commonly used resampling scheme, but alternatives such as stratified resampling give improvements in sample variance (Liu and Chen, 1998; Douc et al., 2005).

For the SDE models considered here, resampling is not sufficient to overcome the degeneracy problem. Our solution is to avoid using the model as the proposal distribution, and to instead build novel Brownian bridge type proposals, based on the proposals developed in Golightly and Wilkinson (2008), that guide the particles toward the next data point (thus decreasing degeneracy). The key is to exploit the Euler-Maruyama approximation we use to simulate from the underlying SDE, in order to condition the proposal distribution on the next observation, thus increasing the number of proposals with large weights. Each of our models are SDEs of the form

$$\mathrm{d}\boldsymbol{X}(t) = \boldsymbol{\mu}(\boldsymbol{X}(t), \boldsymbol{\theta})\,\mathrm{d}t + \Sigma_X^{\frac{1}{2}}(\boldsymbol{X}(t), \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{W}(t).$$

The Euler-Maruyama approximation simulates from the SDE over time interval Δt by partitioning the interval into J sub-intervals of length $\delta t = \frac{\Delta t}{J}$, and using the discrete time equation

$$\boldsymbol{X}(t' + \delta t) = \boldsymbol{\mu}(\boldsymbol{X}(t'), \boldsymbol{\theta})\,\delta t + \Sigma_X^{\frac{1}{2}}(\boldsymbol{X}(t'), \boldsymbol{\theta})\,\delta t^{\frac{1}{2}}\boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t$ is a vector of independent standard Gaussian random variables. Simulating from the discrete time equation between two observation times, τ_m and τ_{m+1}, introduces $(J-1) \times d$ latent variables, $\boldsymbol{X}_{m-1,1}, ..., \boldsymbol{X}_{m-1,J-1}$, where we let $\boldsymbol{X}_{m,j} = \boldsymbol{X}(\tau_m + j \cdot \delta t)$. We can extend the particle filter to also sample from these latent variables, by using a proposal distribution of the form $\tilde{r}_{m+1}(\boldsymbol{X}_{m,1}, ..., \boldsymbol{X}_{m,J} \mid Y_{m+1}, \boldsymbol{X}_m, \boldsymbol{\theta})$. The importance weight calculation in the particle filter is then

$$\omega_{m+1}^{(k)} = \frac{\prod_{j=1}^{J} \pi(\boldsymbol{X}_{m,j} \mid \boldsymbol{X}_{m,j-1}, \boldsymbol{\theta})\, \pi(Y_{m+1} \mid \boldsymbol{X}_{m+1}, \boldsymbol{\theta})}{\tilde{r}_m(\boldsymbol{X}_{m,1}, ..., \boldsymbol{X}_{m,J} \mid Y_{m+1}, \boldsymbol{\theta})},$$

where the $\pi(\boldsymbol{X}_{m,j} \mid \boldsymbol{X}_{m,j-1}, \boldsymbol{\theta})$ are now assumed to be Gaussian densities.

We can guide the particles into regions of high likelihood by conditioning the value of $\boldsymbol{X}_{m,j+1}$ on future observation, Y_{m+1}. This can be done by approximating the distribution of Y_{m+1} conditional on $\boldsymbol{X}_{m,j}$ using a single Euler-Maruyama step of size $\widetilde{\Delta t} = \Delta t - j\delta t$. To do this conditioning, note that under an Euler-Maruyama step of interval size $\widetilde{\Delta t}$,

$$\boldsymbol{X}_{m+1} \mid \boldsymbol{X}_{m,j}, \boldsymbol{\theta} \sim \mathcal{N}_d\left(\boldsymbol{X}_{m,j} + \boldsymbol{\mu}_{m,j}\widetilde{\Delta t},\, \Sigma_{m,j}\widetilde{\Delta t}\right),$$

where $\boldsymbol{\mu}_{m,j} = \boldsymbol{\mu}\left(\boldsymbol{X}_{m,j}, \boldsymbol{\theta}\right)$ and $\Sigma_{m,j} = \Sigma_X\left(\boldsymbol{X}_{m,j}, \boldsymbol{\theta}\right)$. We can then see that the joint distribution of $\boldsymbol{X}_{m,j+1}$ and Y_{m+1}, given $\boldsymbol{X}_{m,j}$, is

$$\begin{pmatrix} \boldsymbol{X}_{m,j+1} \\ Y_{m+1} \end{pmatrix} \mid \boldsymbol{X}_{m,j}, \boldsymbol{\theta} \sim \mathcal{N}_{d+1}\left(\begin{pmatrix} \boldsymbol{X}_{m,j} + \boldsymbol{\mu}_{m,j}\delta t \\ \boldsymbol{H}\left(\boldsymbol{X}_{m,j} + \boldsymbol{\mu}_{m,j}\widetilde{\Delta t}\right) + D \end{pmatrix}, \begin{pmatrix} \Sigma_{m,j}\delta t & \Sigma_{m,j}\boldsymbol{H}^\top \delta t \\ \boldsymbol{H}\Sigma_{m,j}\delta t & \boldsymbol{H}\Sigma_{m,j}\boldsymbol{H}^\top \widetilde{\Delta t} + \sigma_y^2 \end{pmatrix}\right).$$

Conditioning this distribution on Y_{m+1} (Eaton, 1983), then suggests proposals of the form

$$\boldsymbol{X}_{m,j+1} \mid \boldsymbol{X}_{m,j}, Y_{m+1}, \boldsymbol{\theta} \sim \mathcal{N}_d\left(\boldsymbol{\mathcal{M}}_{m,j}, \boldsymbol{\mathcal{S}}_{m,j}\right),$$

where

$$\boldsymbol{\mathcal{M}}_{m,j} = \boldsymbol{X}_{m,j} + \boldsymbol{\mu}_{m,j}\delta t + \boldsymbol{B}^\top \boldsymbol{A}^{-1}\left(Y_{m+1} - \boldsymbol{H}\left(\boldsymbol{X}_{m,j} + \boldsymbol{\mu}_{m,j}\widetilde{\Delta t}\right) - D\right),$$

and

$$\boldsymbol{\mathcal{S}}_{m,j} = \Sigma_{m,j}\delta t - \boldsymbol{B}^\top \boldsymbol{A}^{-1}\boldsymbol{B},$$

with

$$\boldsymbol{A} = \left(\boldsymbol{H}\Sigma_{m,j}\boldsymbol{H}^\top \widetilde{\Delta t} + \sigma_y^2\right) \text{ and } \boldsymbol{B} = \boldsymbol{H}\Sigma_{m,j}\delta t.$$

In our experiments, we have found that using these guided proposals dramatically reduces particle degeneracy. This improves the likelihood estimates, thus increasing the efficiency of the algorithm. Consequently, smaller value of N_x (fewer state particles) can be used, and the PMMH rejuvenation step has better mixing properties, allowing for shorter chains.

3.3. Further details

The tuning parameters are the number of particles, N_θ and N_x, and the proposal distributions for the PMMH rejuvenation steps. Typically, N_θ will be decided by the available computational resource. A low value of N_x can be used for early iterations, but must be increased when using longer time series of data in later iterations. An insufficient number of state particles has a negative impact on the PMCMC acceptance rate, leading to fewer acceptances. This will be reflected in a low particle diversity (the number of unique particles), which needs to be monitored throughout. Automatic calibration of N_x is discussed in Chopin et al. (2013), where it is suggested that N_x is doubled whenever the acceptance rate of the PMCMC step becomes too small. We use $N_x = N_\theta = 1000$ throughout. The fact that we have a collection of particles in each iteration allows automated calibration of the PMMH proposals. For example, using the sample mean and variance to design a random-walk proposal, or using a Gaussian independence sampler. We use independent Gaussian proposals using the sample mean and covariance, with a chain length of 10 to maintain a high particle diversity. In the first iteration, the proposal $r_1\left(\boldsymbol{X}_1 \mid Y_1, \boldsymbol{\theta}\right)$ in the particle filter is taken to be the state prior distributions. In later iterations, for the continuous state variables we again use independent Gaussian proposals using the sample mean and covariance, and for the discrete state variables we use independent Bernoulli proposals where the success probability is empirically estimated using the current sample up to a

maximum (minimum) of 0.95 (0.05). Finally, we check if the algorithm has converged by ensuring that the results are consistent between independent runs.

4. Results

4.1. Simulation study

In order to gain confidence in the ability of our SMC^2 algorithm for both model selection and calibration, we begin with a simulation study. We simulate a single random trajectory from a given model and parameter setting and draw observations from the observation process. We then show that the posterior distributions recover the true value of the parameters (Figure 3) and the true underlying state (Figure 4), and that the Bayes factors correctly identify the true generative model (Table 2). Further simulation studies and details are available in Carson (2015).

We present results from two datasets simulated from SM91: one in which data are from an unforced version, denoted SM91-u, in which parameters $\gamma_P = \gamma_C = \gamma_E = 0$ so that $F = 0$, and a forced version, SM91-f, for which these parameters and F are non-zero. The parameter values used were: $p = 0.8$, $q = 1.6$, $r = 0.6$, $s = 1.4$, $v = 0.3$, $\sigma_1 = 0.2$, $\sigma_2 = 0.3$, $\sigma_3 = 0.3$, $D = 3.8$, $S = 0.8$, $\sigma_y = 0.1$, and additionally for SM91-f $\gamma_P = 0.3$, $\gamma_C = 0.1$, $\gamma_E = 0.4$, which are comparable with those estimated from real data. We simulate observations every 3 kyr over the past 780 kyr to give 261 observations in each dataset, comparable to a low resolution sediment core. From these datasets we calculated the model evidence and posteriors for each of five models: the forced and unforced versions of SM91 and T06, and the forced model PP12. We do not consider an unforced PP12 model because the deglaciation-glaciation transition depends only on the astronomical forcing (whereas SM91 and T06 both oscillate in the absence of any external forcing). The models contain between 10 and 16 parameters. The priors used for each model are given in Table 1.

The estimated $\log_{10}$ Bayes factors ($\log_{10}$ BF) are given in Table 2. A common interpretation suggests that $\log_{10} B_{12} > 0.5$ is substantial evidence in favour of model $\mathcal{M}_1$ over model $\mathcal{M}_2$, $\log_{10} B_{12} > 1$ is strong evidence that $\mathcal{M}_1$ is superior, and $\log_{10} B_{12} > 2$ is very strong evidence (Kass and Raftery, 1995). Conversely, a negative score indicates the same strength of evidence but in the other direction (for $\mathcal{M}_2$ over $\mathcal{M}_1$). In each column, the $\log_{10}$ BF is with respect to the true generative model, so that positive values indicate support for that model over the true model, and negative values indicate support for the true model. Because the $\log_{10}$ BF is just the difference between the log evidences, we can reconstruct the evidences by noting that the $\log_{10}$ evidence ($\log_{10} \pi(y_{1:M}|\mathcal{M})$) is 29.8 for the unforced version of SM91 on the SM91-u dataset, and 40.7 for the forced SM91 model on the SM91-f dataset.

Table 2 about here.

For both simulated datasets, we find a strong preference for the correct model. When applied to

SM91-f, the correct model (the forced SM91 model) is overwhelmingly favoured. The $\log_{10}$ BF to the next most supported model (the forced T06 model) is estimated to be 10.2, indicating decisive evidence in favour of the true model. It is interesting to note that if we remove the forced SM91 model from the analysis, we find decisive evidence in favour of the forced T06 model over any of the other unforced models (a $\log_{10}$ BF of at least 11), showing that the astronomical forcing has explanatory power even in the wrong model (to find other BFs, note that $\log B_{ij} = \log B_{i0} - \log B_{j0}$). This is not particularly surprising, because in both models the astronomical forcing acts as a synchronisation agent, controlling the timing of terminations, and has a strong effect on the likelihood. This is a reassuring finding: it suggests that paleoclimate scientists can implicitly rely upon this effect when arguing for the importance of the astronomical forcing, as it allows us to infer its importance even when using an incorrect model (for we surely are).

When applied to SM91-u the log BF again correctly identifies the correct generative model, although the support for the unforced and forced SM91 models is now much closer (with a $\log_{10}$ BF of 2.1 in favour of the unforced model). In cases where the forcing does not add any explanatory power this is an expected result, as the unforced version of SM91 is nested within the forced version, and can be recovered by setting $\gamma_P = \gamma_C = \gamma_E = 0$. This effect is also noticeable when comparing the forced and unforced T06 models, with the unforced version being preferred with a $\log_{10}$ BF of 1.8.

The marginal posterior distributions for the parameters for the forced SM91 model applied to the SM91-f dataset are shown in Figure 3. We are able to recover the parameters used to generate the data, with the true values lying in regions of high posterior probability. The posteriors for q and σ_3 do not deviate much from the prior, suggesting that a wide range of values explain the data equally well.

Figure 3 about here.

Finally, the sequence of 95% highest density regions (HDRs) for the state estimates for the forced SM91 model applied to the SM91-f dataset are shown in Figure 4. For each of the three states, most of the true values lie within the HDRs, demonstrating that we are able to recover the state of the system, despite only having noisy observations of a single state.

Figure 4 about here.

4.2. ODP677

We now analyse dataset ODP677 from the ocean drilling programme (ODP). The estimated $\log_{10}$ BFs for each model are given in Table 3. We use ODP677-u to refer to age model estimates derived by Huybers (2007) using a depth derived model, and ODP677-f to the astronomically tuned age model estimates described in Lisiecki and Raymo (2005). The BFs are given in comparison to the best model for each dataset. For ODP677-u, the unforced T06 model is best supported with a $\log_{10}$

evidence of 28.2. The results suggest strong evidence in favour of the unforced models over the forced models, which are penalised for containing extra parameters with little explanatory power. That the unforced model is preferred may be surprising compared to earlier works based on similar records (Raymo, 1997; Huybers, 2011); this is discussed further in the conclusions.

Table 3 about here.

When we analyse ODP677-f, the astronomically tuned data, the results are reversed. We now find strong evidence in favour of the PP12 model (with a $\log_{10}$ evidence of 33.7), and that the three forced models are all decisively preferred to the two unforced models, i.e., we find overwhelming evidence using these data that astronomical forcing is necessary to explain the data. The orbital tuning of ODP677-f is the most likely explanation for this. There is some evidence that T06 is more strongly supported than SM91 with a $\log_{10}$ BF of 0.8.

This result is our second key finding. Namely, that inference about the best model is affected by the age model used to date the data. It is vital that modelling assumptions in the dating methods should be understood when performing inference on paleoclimate data. Given that the two chronologies, ODP677-f and ODP677-u, are considered consistent once we account for dating uncertainties, we suggest that this formally demonstrates that the approach of first dating the data, and then carrying out down-stream analyses given this dating (ignoring the uncertainty) may undermine any subsequent inference about the dynamic mechanisms at play.

Whilst these results demonstrate that the age model strongly influences the conclusions in model comparison experiments, care must be taken to not over-interpret the results. Bayes factors are sensitive to changes in the prior distributions, and in this case, are subject to Monte Carlo error. It is possible that selecting new prior distributions consistent with our approach in Section 2.5, or even keeping the same prior distributions and reinitialising the algorithm with a different seed could result in different conclusions. This is investigated in the next section.

The marginal posterior distributions of the parameters in the SM91 model when fit to the ODP677-f data are shown in Figure 5. The astronomical forcing scaling parameters γ_P and γ_E have very small posterior support at 0, suggesting that both precession and obliquity are important.

Figure 5 about here.

Figure 6 provides the density of the ratios $\frac{\sqrt{\gamma_P^2+\gamma_C^2}}{\gamma_E}$, and the argument of the complex number $\gamma_P + i\gamma_C$ for SM91 and T06. PP12 is omitted as the truncation of the forcing makes the parameters incomparable. The rationale for using $\arg(\gamma_P + i\gamma_C)$ can be made clear by noting that the forcing (Equation 1) can be reformulated as

$$F(t;\boldsymbol{\gamma}) = \Gamma \sin(\varpi(t) + \phi) + \gamma_E E(t),$$

where $\Gamma \propto |\gamma_P + i\gamma_C|$ and $\phi = \arg(\gamma_P + i\gamma_C)$ (the proportionality is straightforwardly set by the normalisation constants). Hence, Γ controls the amplitude of the effects associated with the physical

process of precession of equinoxes, while ϕ controls the phase. It follows that the ratio $\frac{\Gamma}{\gamma_E} = \frac{\sqrt{\gamma_P^2+\gamma_C^2}}{\gamma_E}$ measures the relative weight in the forcing of two different physical effects: the precession of equinoxes, and the changes in the tilt of the Earth's equator (obliquity). The results in the left plot of Figure 6 are consistent across models, with a ratio lower than one, suggesting that the control of obliquity dominates. Translated in terms of paleoclimate dynamics, this means that ice age dynamics are controlled by insolation integrated over a season length, rather that just the maximum insolation over the year. The right plot in Figure 6 shows the argument of the complex number $\gamma_P + i\gamma_C$. Here, zero means that phase of the precession forcing matches that of the June solstice insolation. A phase of $\pi/2$ would mean that the system is controlled by March insolation, while $-\pi/2$ would point to September insolation. All densities are broadly centred on zero, suggesting a summer insolation control in the Northern Hemisphere (this is also consistent with a winter control in the Southern Hemisphere, but this is less plausible). The nominal uncertainty of approximately 0.8 radians translates into an uncertainty of about 2 months in calendar time. Physically, it is reasonable to assume that the driving effect changes as ice sheets grow and melt, and that these changes contribute to the variance of the density curves. We acknowledge that dating assumptions will also presumably be crucial in determining this quantity.

Figure 6 about here.

4.3. Robustness

The simulation studies show that there is sufficient information in the observations to easily detect the correct parametric form of the model in each case. However, the Bayes factors are sensitive to both Monte Carlo error in the evidence estimates, the choice of prior distributions, and model misspecification, and so care needs to be taken when using real data in order to avoid over-interpretation.

To investigate the Monte Carlo error in the Bayes factor estimates, we calculate the model evidences for the forced models on ODP677-f using ten randomly generated seeds for each model. The $\log_{10}$ evidences were in the range $[28.4, 29.0]$ for SM91, $[29.4, 31.5]$ for T06, and $[33.3, 34.2]$ for PP12. Over all repeats PP12 is still strongly favoured, but not necessarily decisively favoured (as the $\log_{10}$ BF is as small as 1.8). Likewise, T06 is favoured over SM91, but not necessarily substantially favoured (as the $\log_{10}$ BF is as small as 0.4). Forced models are still decisively favoured over unforced models. The implications for the experiment on ODP677-u are that each of the forced models are plausibly equally well supported by the data, and further confirming that the difference between the forced and unforced version of the same parametric model is small. The magnitude of the estimation error can be decreased by using more particles in the SMC^2 algorithm, but this will require very long computational runs. Using $N_x = N_\theta = 1000$ takes 3-4 days on a standard desktop depending on the model. However, SMC algorithms are well suited to run in parallel, and we were able to obtain a $\sim$25$\times$ speed-up on a Tesla K20 GPU.

To test the robustness of our conclusions to changes in the prior distributions, we performed a sensitivity analysis using the two ODP-677 datasets. Kass and Raftery (1995) advocate recomputing the Bayes factors with perturbed hyperparameters, for example by halving/doubling scale and variance parameters. Given the computational expense of our experiments, it is not feasible to do this one parameter at a time. Instead we conducted the following two experiments for the forced models. In the first we maintained the 1:2:5 cross-model rules for the forcing prior construction (as discussed in Section 2.5), but each of the other hyperparameters we halve or double with equal probability. For the gamma distributions we maintain the prior means by doubling or halving the shape parameter (exponential distributions are treated as gamma distributions with a shape parameter of 1 for this purpose). In the second experiment we halve the hyperparameters for the astronomical forcing parameters only, and compare the results to those obtained in the initial experiment in order to violate the 1:2:5 ratio. The aim of the first experiment is to investigate the sensitivity of our conclusions to changes in the prior distributions for the model parameters, whereas the aim of the second is to test the sensitivity to the cross-model assumptions made in Section 2.5.

Table 4 about here.

The estimated $\log_{10}$ BFs of the new experiments are shown in Table 4, given with respect to the best model from the first analysis for each dataset. We have denoted results from the first sensitivity analysis as MP, to indicate that our focus is on the model parameters, and results from the second sensitivity analysis as FP, as our focus is on the forcing parameters. It is difficult to decouple the effect of changing the prior distributions from the Monte Carlo error, but the combined change is within ± 2 in each case. Taken together with the suggested interpretation of BFs in Kass and Raftery (1995), this suggests the conservative rule-of-thumb that the $\log_{10}$ BF should be at least 5 to constitute strong evidence for one model over another if we are not confident about our choice of prior distributions. If we revisit the results from the first analysis with this in mind, we find broadly the same model comparison conclusions, i.e., unforced T06 is favoured in ODP677-u, and PP12 is favoured in ODP677-f, even when the prior scaling rules designed in Section 2.5 are violated, but we can no longer be confident that these results constitute strong evidence in favour of these models. We still have strong evidence against unforced models for ODP677-f.

As our primary focus is on selecting between phenomenological models, we also need to consider the impact of the modelling assumptions used for the stochastic driver and observation errors. In particular, if these components are mis-specified, what will the effect be on the Bayes factors? To test the robustness of the Bayes factors we revisit the simulation study datasets from Section 4.1 and simulate additional sets of observations as follows. Firstly, we simply redraw observations ($Y_{1:T}$) from the trajectory generated for SM91-f ($\boldsymbol{X}_{1:T}$) in order to give a baseline for the Bayes factor variability under a redrawing of the observation errors. We term this dataset SS-2. We then draw observations,

again from the same trajectory, using the AR1 process

$$\epsilon_{t+1} = \rho\epsilon_t + \mathcal{N}\left(0, \sigma^2\right),$$

with $\sigma = 0.1$, and $\rho = 0.6$ (termed SS-AR06) and $\rho = 0.9$ (termed SS-AR09). The purpose of this is to test the robustness of the Bayes factors under a mis-specification of the observation error model. We also generate two new trajectories from SM91-f, SS-JD78 and SS-JD780, using a jump-diffusion instead of a Brownian motion as the stochastic driver of the SDE. The jump times are drawn from a Poisson process with intensity 0.1 kyr^{-1} in SS-JD78, and 1 kyr^{-1} in SS-JD780, giving the expected number of jumps as 78 and 780 respectively. The jump intensity is multivariate Gaussian with mean 0 and covariance matrix $\text{diag}(0.2^2, 0.3^2, 0.3^2)$. This is 100 times the variance of a single Euler-Maruyama integration step of the Brownian motion, and equivalent to about 7-10% of the range of each state variable. This allows us to test the robustness of the Bayes factors under a mis-specification of the stochastic driver. Finally, we generate two datasets where we expect the Bayes factors to be weak, in order to ensure that the results are not overconfident. In the first, termed SS-LE, we generate observations using the same trajectory as SM91-f, but with large observation errors (increasing the variance of the error 100-fold). In the second, termed SS-OU, we generate a trajectory using an Ornstein-Uhlenbeck process (i.e., not using any of the phenomenological models). The datasets are shown in Figure 7.

Figure 7 about here.

The estimated $\log_{10}$ Bayes factors (with respect to the best model) for our five models in each of the seven new datasets are given in Table 5. We emphasise that care needs to be taken when interpreting Bayes factors between different datasets: Bayes factors only indicate relative model performance, and so we may obtain larger Bayes factors in instances where the correct model is not in our library. This should not be interpreted as the preferred model being good in an absolute sense. For SS-2, the model evidences change by a factor of 10^6 (cf. Table 2) and the Bayes factors change by up to 2.6 on the $\log_{10}$ scale (some of this variability will be due to the Monte Carlo error), but our conclusions are fundamentally unchanged. The exception is that the unforced SM91 model is now preferred over PP12, but the Bayes factor between the two models is relatively small in both simulation studies. In the AR1 datasets, T06 seems to perform poorly (the Bayes Factor between SM91 and PP12 seems to be consistent with the original simulation study). A potential reason for this is that the interglacial-glacial switches in T06 occur at some constant threshold for the state-variable, whereas in the data there is greater variation in the local minima (particularly in SS-AR09). SM91 and PP12 seem to be more robust to this behaviour. For SS-JD78, the Bayes factors seem consistent with the initial simulation study. This can partially be explained by the nature of oscillators such as SM91, where trajectories following perturbations converge rapidly to a limit cycle, and so the overall trajectory might be robust to large jumps. However, we can see in Figure 7 that the cycles seem

less predictable, indicating that the jumps have had some effect. For SS-JD780 the Bayes factors are much weaker, and we can see in Figure 7 that the cycles are far less regular than the SM91 model forced by Brownian motion. However, we still favour the forced SM91 model, suggesting that enough information is preserved to allow us to pick out the correct model. For SS-LE (large observation error) the order of the models is consistent with the previous simulation studies, but the Bayes factors are much weaker, as expected. For SS-OU (data generated from an Ornstein-Uhlenbeck process), none of our five candidate models are close to being useful, and consequently we find relatively small BFs compared to previous simulation studies, but with the models ordered by their complexity (in terms of the number of parameters), demonstrating Occam's razor type behaviour of the Bayes factor.

Table 5 about here.

These are reassuring findings, in that the order of the models seem to be robust to moderate mis-specifications of the stochastic driver and observation error model. In the limit where no model has explanatory power, the Bayes factors favour the less complex models. It is plausible that significant model mis-specification could bias the model selection to a significant extent, e.g., by favouring more complex models to account for the mis-specification. Ideally alternative drivers and observation error models would be considered, with Bayes factors used to select between candidates. However, each would require the design of new proposal mechanisms in the particle filter, which is a non-trivial task.

5. Conclusions

We have two key conclusions. The first is that Monte Carlo methodology and computer power are now sufficiently advanced that with work, it is possible to fully solve the Bayesian model selection problem for a wide class of phenomenological models of the glacial-interglacial cycle. Using only 261 observations, we are able to learn up to 16 parameters, state trajectories containing 261×3 values, and calculate the model evidence. Moreover, these evidences are sufficiently different (and able to be estimated with sufficient accuracy) that we can discriminate between the ability of the models to explain the data. However, care needs to be taken so as to not over-interpret the results from these experiments. Firstly, Bayes factors are known to be sensitive to the prior distributions, and so the conclusions might not be robust to changes in the priors. Fortunately, our analysis indicates that useful results can still be obtained when the prior distributions are carefully elicited. Secondly, that one dynamical system is more supported by the data than another does not necessarily imply that the *physical* interpretation of that model is valid. At this level of conceptual modelling, different physical interpretations may produce similar equations. This point has been made before (Tziperman et al., 2006) and we add here that the stochastic differential equations emerge as a combination of judgements on physical processes *and* model discrepancy, embedded in the stochastic parameterisations. On the other hand, the Bayesian formalism for choosing between models offers a natural starting point for

developing a physical interpretation, and knowledge of physical constraints can be incorporated within the parameter prior distributions. Physical disambiguation will also arise as the complexity of the model is increased, and as more diverse datasets are used in the inference process.

Our second conclusion concerns the need to avoid "theory-laden" data. The results from analysing the ODP677 data, show that the age model used to date the core become critical when the data are subsequently used to make scientific judgements. The astronomically-tuned age model gives support for a model in which ice ages are *driven* by the astronomical forcing (that is, without an underlying autonomous limit cycle), while the age model which was not tuned on astronomical forcing favours models explaining ice ages as an autonomous limit cycle. These are two qualitatively different explanations of ice ages. Admittedly, fifty years of climate research have established beyond doubt that the astronomical forcing affects the climate system enough to interfere with ice ages dynamics: the rejection of astronomical forcing here must presumably be explained by errors in the ODP677-u time scale. On the other hand, we observed that both time scales were compatible with uncertainties provided by the respective authors. This suggests that analysing the data in stages, cutting feedbacks between uncertainties, does not only affect the conclusion about the role of the astronomical forcing, it also affects inferences about the internal system dynamics. We believe that this is the first time the effect of the age model on subsequent analyses has been so clearly demonstrated. Given the coupling between the model parameters and the age estimates, instead of first dating the core and then using those dates (with or without uncertainties), we need to jointly estimate the age model at the same time as testing further hypotheses, accounting for all the joint uncertainties (Carson, 2015).

The experiments included in this paper can be extended in several ways. Firstly, we considered only a handful of models, and both the number and complexity of models can be increased. With the approach described here, extra models can be included by running the SMC^2 algorithm for each model. This has the benefit that the entire experiment does not need to be redesigned/repeated for different combinations of models. Different astronomical forcings can also be considered. For example, the astronomical forcing terms are often tested independently. This can easily be achieved by setting undesired astronomical scaling terms to 0 in our forced models. Making the forcing term state dependent, so that an increase in sea-ice increases albedo, which in turn alters the influence of variation in insolation, is also a possibility.

Finally, we do not need to limit ourselves to a single dataset. The observation model can be extended to compare the state of the system to multiple cores. Likewise, multivariate observations could be used; SM91 models both ice volume and CO_2 concentration, and records exist for both of these quantities.

Overall, we hope that this work acts as a proof of concept. Careful statistical analysis combining data and models can lead to insights in paleoclimate science.

6. Supplementary Information

The programs and data that were used in this article can be obtained from http://wileyonlinelibrary.com/journal/rss-datasets.

References

Abe-Ouchi, A., Saito, F., Kawamura, K., Raymo, M. E., Okuno, J., Takahashi, K. and Blatter, H. (2013) Insolation-driven 100,000-year glacial cycles and hysteresis of ice-sheet volume. *Nature*, **500**, 190–193.

Ando, T. and Tsay, R. (2010) Predictive likelihood for bayesian model selection and averaging. *International Journal of Forecasting*, **26**, 744–763.

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, **72**, 269–342.

Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, **37**, 697–725.

Ashkenazy, Y. and Tziperman, E. (2004) Are the 41 kyr glacial oscillations a linear response to Milankovitch forcing? *Quaternary Science Reviews*, **23**, 1879–1890.

Berger, A. (1978) Long term variations of daily insolation and Quaternary climate changes. *Journal of Atmospheric Sciences*, **35**, 2362–2367.

Berger, A. and Loutre, M. (2004) Astronomical theory of climate change. *Journal de Physique IV*, **121**, 1–35.

Brovkin, V., Ganopolski, A., Archer, D. and Munhoven, G. (2012) Glacial co_2 cycle as a succession of key physical and biogeochemical processes. *Climate of the Past*, **8**, 251–264. URL`http://www.clim-past.net/8/251/2012/`.

Carson, J. (2015) *Uncertainty Quantification in Palaeoclimate Reconstruction*. Ph.D. thesis, University of Nottingham.

Chopin, N., Jacob, P. E. and Papaspiliopoulos, O. (2013) SMC^2: an efficient algorithm for sequential analysis of state-space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 397–426.

Crucifix, M. (2011) How can a glacial inception be predicted? *The Holocene*, **21**, 831–842.

— (2012) Oscillators and relaxation phenomena in Pleistocene climate theory. *Transactions of the Philosophical Transactions of the Royal Society A*, **370**, 1140–1165.

— (2013) Why could ice ages be unpredictable. *Climate of the Past*, **9**, 2253–2267.

Del Moral, P. (2004) *Feynman-Kac Formulae*. Springer.

Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *Journal of the Royal Society Series B*, **68**, 411–436.

Douc, R., Cappé, O. and Moulines, E. (2005) Comparison of resampling schemes for particle filtering. In *4th International Symposium on Image and Signal Processing and Analysis (ISPA)*.

Doucet, A. and Johansen, A. M. (2009) A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, **12**, 656–704.

Eaton, M. L. (1983) *Multivariate Statistics: a vector space approach*. John Wiley and Sons.

Elderfield, H., Ferretti, P., Greaves, M., Crowhurst, S., McCave, I. N., Hodell, D. and Piotrowski, A. M. (2012) Evolution of ocean temperature and ice volume through the mid-pleistocene climate transition. *Science*, **337**, 704–709.

Emiliani, C. (1955) Pleistocene temperatures. *Journal of Geology*, **63**, 538–578.

Feng, F. and Bailer-Jones, C. A. L. (2015) Obliquity and precession as pacemakers of Pleistocene deglaciations. *Quaternary Science Reviews*, **122**, 166–179.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for bayesian models. *Statistics and Computing*, **24**, 997–1016.

Golightly, A. and Wilkinson, D. (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statististics & Data Analysis*, **52**, 1674–1693.

Gordon, N., Salmond, D. and Smith, A. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F*, **140**, 107–113.

Huybers, P. (2007) Glacial variability over the last two million years: an extended depth-derived age-model, continuous obliquity pacing, and the pleistocene progression. *Quaternary Science Reviews*, **26**, 37–55.

— (2011) Combined obliquity and precession pacing of late Pleistocene deglaciations. *Nature*, **480**, 229–232.

Huybers, P. and Wunsch, C. (2005) Oliquity pacing of late Pleistocene terminations. *Nature*, **434**, 491–494.

Imbrie, J. and Imbrie, J. Z. (1980) Modelling the climatic response to orbital variations. *Science*, **207**, 943–953.

Imbrie, J. J., Hays, J. D., Martinson, D. G., McIntyre, A., Mix, A. C., Morley, J. J., Pisias, N. G., Prell, W. L. and Shackleton, N. J. (1984) The orbital theory of Pleistocene climate: Support from a revised chronology of the marine δO^{18}record. In *Milankovitch and Climate, Part I* (eds. A. Berger, J. Imbrie, J. Hays, J. Kukla and B. Saltzman), 269–305. Norwell, Mass.: D. Reidel.

Imbrie, J. Z., Imbrie-Moore, A. and Lisiecki, L. E. (2011) A phase-space model for pleistocene ice volume. *Earth and Planetary Science Letters*, **307**, 94–102.

Jeffreys, H. (1939) *The theory of probability*. Oxford University Press.

Kass, R. and Raftery, A. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kwasniok, F. (2013) Analysis and modelling of glacial climate transitions using simple dynamical systems. *Philosophical Transactions of the Royal Society A*, **371**, 2011.0472.

Laskar, J., Robutel, P., Joutel, F., Boudin, F., Gastineau, M., Correia, A. C. M. and Levrard, B. (2004) A long-term numerical solution for the insolation quantities of the Earth. *Astronomy and Astrophysics*, **428**, 261–285.

Lisiecki, L. (2010) Links between eccentricity forcing and the 100,000-year glacial cycle. *Nature Geoscience*, **3**, 349–352.

Lisiecki, L. and Raymo, M. (2005) A pliocene-pleistocene stack of 57 globally distributed benthic δ^{18}O records. *Paleoceanography*, **20**, PA1003.

Liu, J. and Chen, R. (1998) Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, **93**, 1032–1044.

Milankovitch, M. (1941) *Kanon der Erdbestrahlung und Seine Anwendung auf das Eiszeitenproblem (Canon of insolation and the ice-age problem)*. Belgrad: Könlishe Serbische Akademie.

— (1998) *Canon of insolation and the ice-age problem*. Beograd: Narodna biblioteka Srbije.

Mitsui, T. and Aihara, K. (2014) Dynamics between order and chaos in conceptual models of glacial cycles. *Climate Dynamics*, **42**, 3087–3099.

Mitsui, T. and Crucifix, M. (2016) Effects of additive noise on the stability of glacial cycles. In *Mathematical Paradigms of Climate Science* (eds. F. Ancona, P. Cannarsa, C. Jones and A. Portaluri), Spring INdAM Series, 93–113. Springer Verlag.

Paillard, D. (1998) The timing of Pleistocene glaciations from a simple multiple-state climate model. *Nature*, **391**, 378–381.

Parrenin, F. and Paillard, D. (2012) Terminations VI and VIII (∼530 and ∼720 kyr BP) tell us the importance of obliquity and precession in the triggering of deglaciations. *Climate of the Past*, **8**, 2031–2037.

Raymo, M. E. (1997) The timing of major climate terminations. *Paleoceanography*, **12**, 577–585.

Roe, G. and Allen, M. (1999) A comparison of competing explanations for the 100,000-yr ice age cycle. *Geophysical Research Letters*, **26**, 2259–2262.

Ruddiman, W. F. (2006) Ice-driven CO_2 feedback on ice volume. *Climate of the Past*, **2**, 43–55.

Saltzman, B. (1988) Modelling the slow climate attractor. In *Physically-Based Modelling and Simulation of Climate and Climatic Change*, 737–754. Springer.

Saltzman, B. and Maasch, K. (1990) A first-order global model of late Cenozoic climate. *Transactions of the Royal Society of Edinburgh: Earth Sciences*, **81**, 315–325.

— (1991) A first-order global model of late Cenozoic climate. II further analysis based on simplification of the CO_2 dynamics. *Climate Dynamics*, **5**, 201–210.

Shackleton, N., Berger, A. and Peltier, W. (1990) An alternative astronomical calibration of the lower Pleistocene timescale based on ODP site 677. *Transactions of the Royal Society of Edinburgh: Earth Sciences*, **81**, 251–261.

Shackleton, N. J. (1967) Oxygen isotope analyses and Pleistocene temperatures re-assessed. *Nature*, **215**.

Shackleton, N. J., Backman, J., Zimmerman, H., Kent, D. V., Hall, M. A., Roberts, D. G., Schnitker, D., Baldauf, J. G., Desprairies, A., Homrighausen, R., Huddlestun, P., Keene, J. B., Kaltenback, A. J., Krumsiek, K. A. O., Morton, A. C., Murray, J. W. and Westberg-Smith, J. (1984) Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the north atlantic region. *Nature*, **307**, 620–623.

Tziperman, E., Raymo, M., Huybers, P. and Wunsch, C. (2006) Consequences of pacing the Pleistocene 100 kyr ice ages by nonlinear phase locking to Milankovitch forcing. *Paleoceanography*, **21**, PA4206.

Vehtari, A. and Ojanen, J. (2012) A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6**, 142–228.

Wan, E., Van Der Merwe, R. et al. (2000) The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, 153–158. IEEE.

Table 1. Prior distributions used for each model in both the simulation study and the analysis of ODP677.

SM91	T06	PP12
$\gamma_P \sim \text{Exp}(1/0.3)$	$\gamma_P \sim \text{Exp}(1/0.6)$	$\gamma_P \sim \text{Exp}(1/1.5)$
$\gamma_C \sim \mathcal{N}(0, 0.3^2)$	$\gamma_C \sim \mathcal{N}(0, 0.6^2)$	$\gamma_C \sim \mathcal{N}(0, 1.5^2)$
$\gamma_E \sim \text{Exp}(1/0.3)$	$\gamma_E \sim \text{Exp}(1/0.6)$	$\gamma_E \sim \text{Exp}(1/1.5)$
$p \sim \Gamma(2, 1/1.2)$	$p_0 \sim \text{Exp}(1/0.3)$	$a \sim \Gamma(8, 0.1)$
$q \sim \Gamma(7, 1/3)$	$K \sim \text{Exp}(1/0.1)$	$a_d \sim \text{Exp}(1)$
$r \sim \Gamma(2, 1/1.2)$	$s \sim \text{Exp}(1/0.3)$	$a_g \sim \text{Exp}(1)$
$s \sim \mathcal{N}\left(0, 1.5^2\right)$	$\alpha \sim \text{Beta}(40, 30)$	$\kappa_P \sim \text{Exp}(1/20)$
$v \sim \Gamma(2, 1/0.3)$	$x_l \sim \text{Exp}(1/3)$	$\kappa_C \sim \mathcal{N}(0, 20^2)$
$\sigma_1 \sim \text{Exp}(1/0.3)$	$x_u \sim \Gamma(90, 0.5)$	$\kappa_E \sim \text{Exp}(1/20)$
$\sigma_2 \sim \text{Exp}(1/0.3)$	$\sigma_1 \sim \text{Exp}(1/2)$	$\tau \sim \text{Exp}(1/10)$
$\sigma_3 \sim \text{Exp}(1/0.3)$		$v_l \sim \text{Exp}(1/5)$
		$v_u \sim \Gamma(220, 0.5)$
		$\sigma_1 \sim \text{Exp}(1/5)$
$D \sim \text{U}(3, 5)$	$D \sim \text{U}(2.5, 4.5)$	$D \sim \text{U}(2.5, 4.5)$
$S \sim \text{U}(0.25, 1.25)$	$S \sim \text{U}(0.02, 0.05)$	$S \sim \text{U}(0.01, 0.03)$
$\sigma_y \sim \text{Exp}(1/0.1)$	$\sigma_y \sim \text{Exp}(1/0.1)$	$\sigma_y \sim \text{Exp}(1/0.1)$
$X_{(1)}(\tau_1) \sim \text{U}(-1.5, 1.5)$	$X_{(1)}(\tau_1) \sim \text{U}(3, 45)$	$X_{(1)}(\tau_1) \sim \text{U}(0, 120)$
$X_{(2)}(\tau_1) \sim \text{U}(-1.5, 1.5)$	$X_{(2)}(\tau_1) \sim \text{Bernoulli}(0.5)$	$X_{(2)}(\tau_1) \sim \text{Bernoulli}(0.5)$
$X_{(3)}(\tau_1) \sim \text{U}(-1.5, 1.5)$		

Table 2. Log Bayes factors for comparing five different models on the two simulated datasets. SM91-u is data generated from an unforced version of SM91, whereas SM91-f is generated from an astronomically forced version of SM91.

Model		Dataset	
		SM91-u	SM91-f
SM91	Forced	-2.1	0
	Unforced	0	-22.2
T06	Forced	-9.5	-10.2
	Unforced	-7.7	-26.0
PP12	Forced	-20.4	-21.2

Table 3. Log Bayes factors for comparing five different models on ODP677. ODP677-u uses observation ages from the non-orbitally tuned H07 stack, and ODP677-f takes observation ages from the orbitally tuned LR04 stack.

Model		Dataset	
		ODP677-u	ODP677-f
SM91	Forced	-3.6	-4.8
	Unforced	-2.2	-16.1
T06	Forced	-2.9	-4.0
	Unforced	0	-12.2
PP12	Forced	-3.8	0

Table 4. Log Bayes factors for the sensitivity analysis on ODP677 compared to the best model in Table 3 (forced models only).

Model		Dataset	
		ODP677-u	ODP677-f
SM91 Forced	MP	−4.1	−5.7
	FP	−2.8	−5.0
T06 Forced	MP	−2.3	−3.4
	FP	−1.5	−2.9
PP12 Forced	MP	−5.5	−1.2
	FP	−3.0	1.4

Table 5. Log Bayes factors for comparing five different models on seven simulated datasets.

Model		Dataset						
		SS-2	SS-AR06	SS-AR09	SS-JD78	SS-JD780	SS-LE	SS-OU
SM91	Forced	0	0	0	0	0	0	−1.4
	Unforced	−22.0	−25.6	−20.0	−24.0	−7.2	−3.2	−0.5
T06	Forced	−12.1	−14.8	−16.9	−12.1	−3.2	−2.5	−1.1
	Unforced	−28.4	−32.3	−33.9	−28.0	−9.7	−4.0	0
PP12	Forced	−23.8	−22.9	−20.1	−22.1	−3.9	−3.0	−1.7

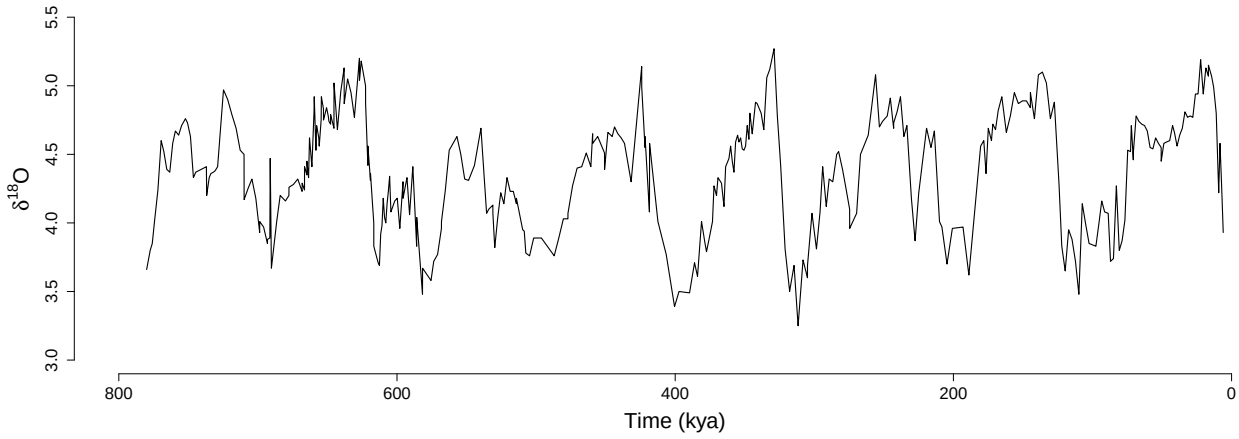

Fig. 1. Observed δ^{18}O from ODP677 (Shackleton et al., 1990) corresponding to the past 780 kyr with the H07 chronology (Huybers, 2007).

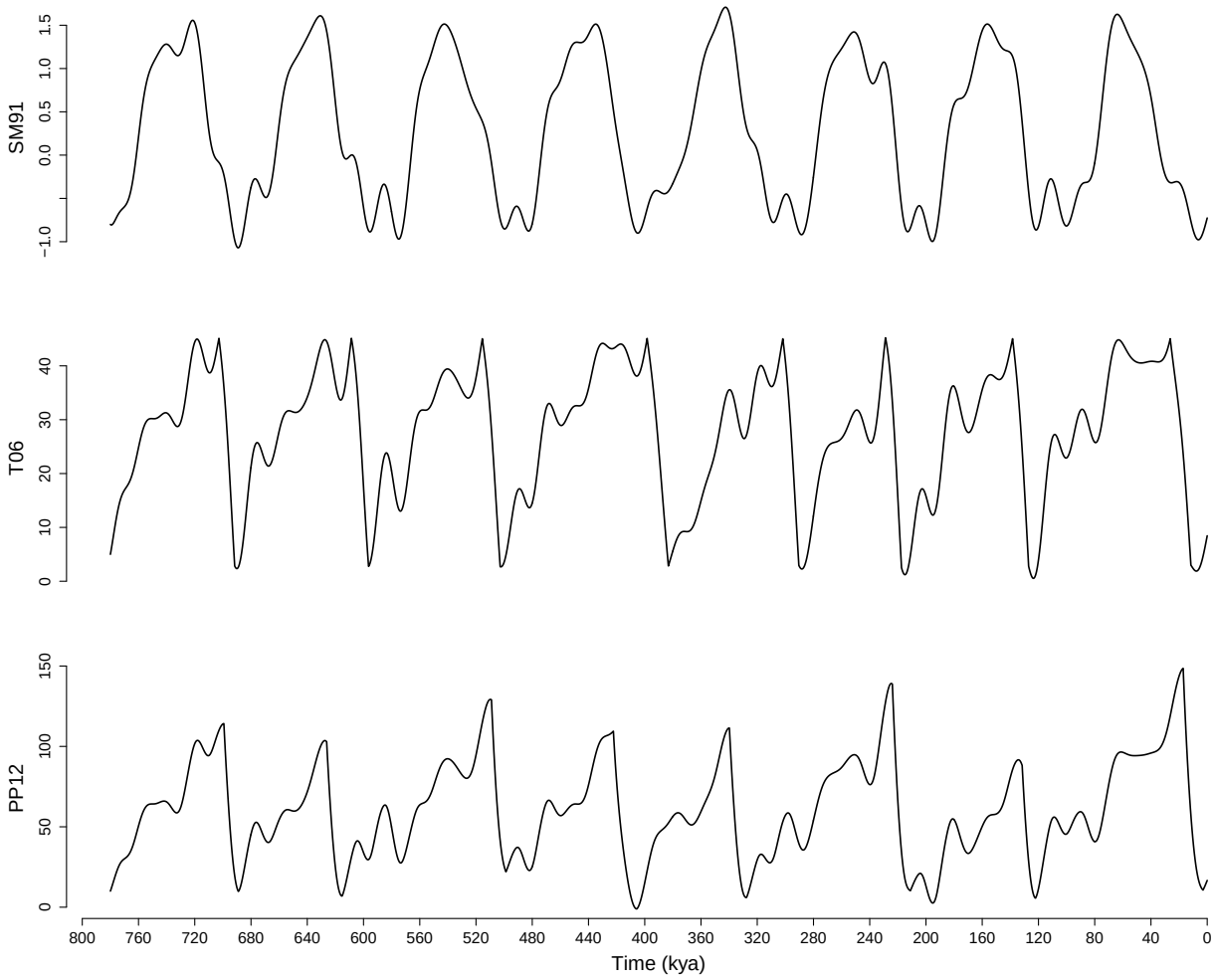

Fig. 2. Comparison of the ice volume generated from deterministic versions of SM91, T06, and PP12.

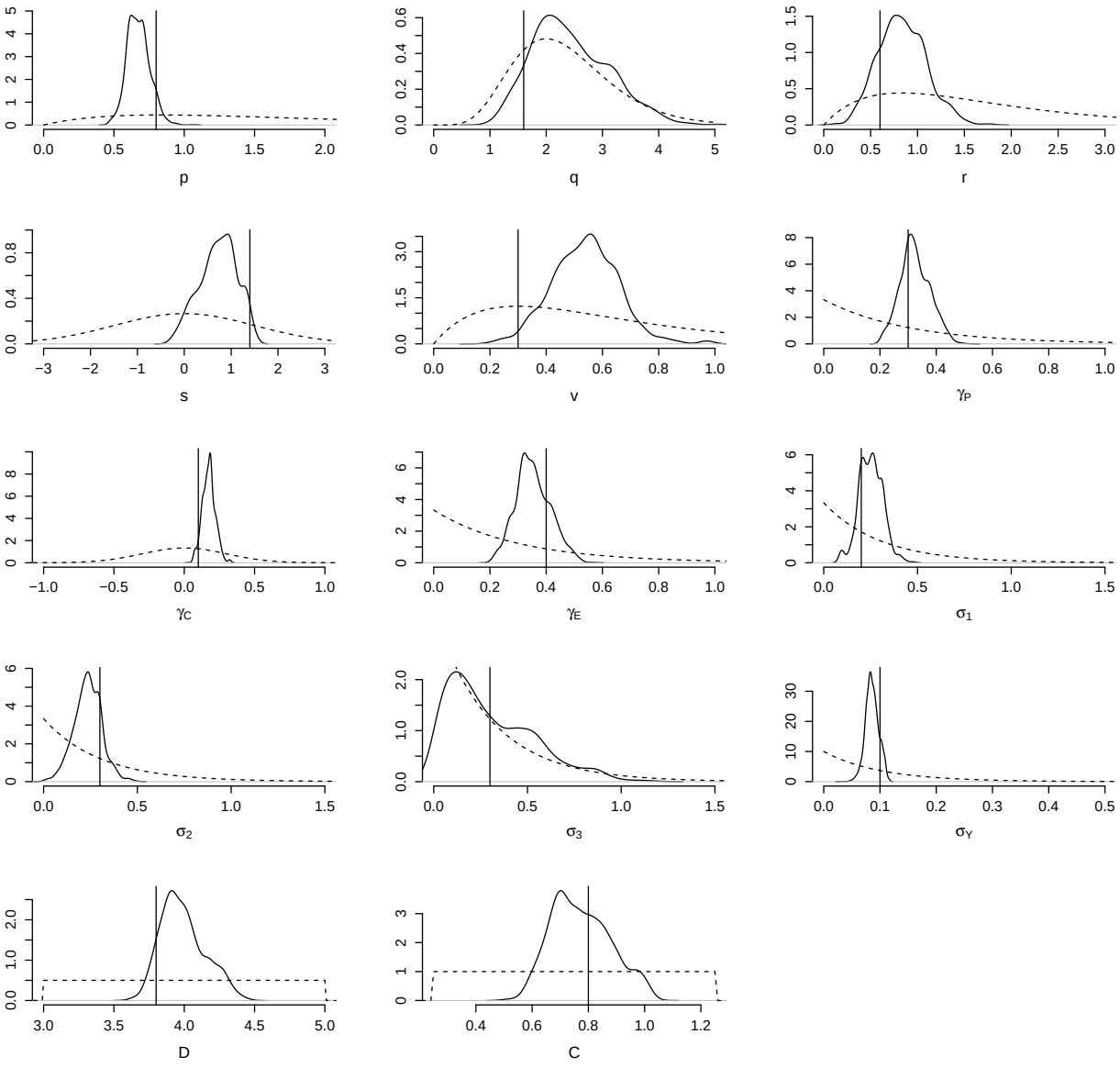

Fig. 3. Marginal posterior distributions for the parameters of the forced SM91 model when fit to the SM91-f dataset. Vertical lines show the parameter values used to generate the data, and dashed lines represent the prior distribution.

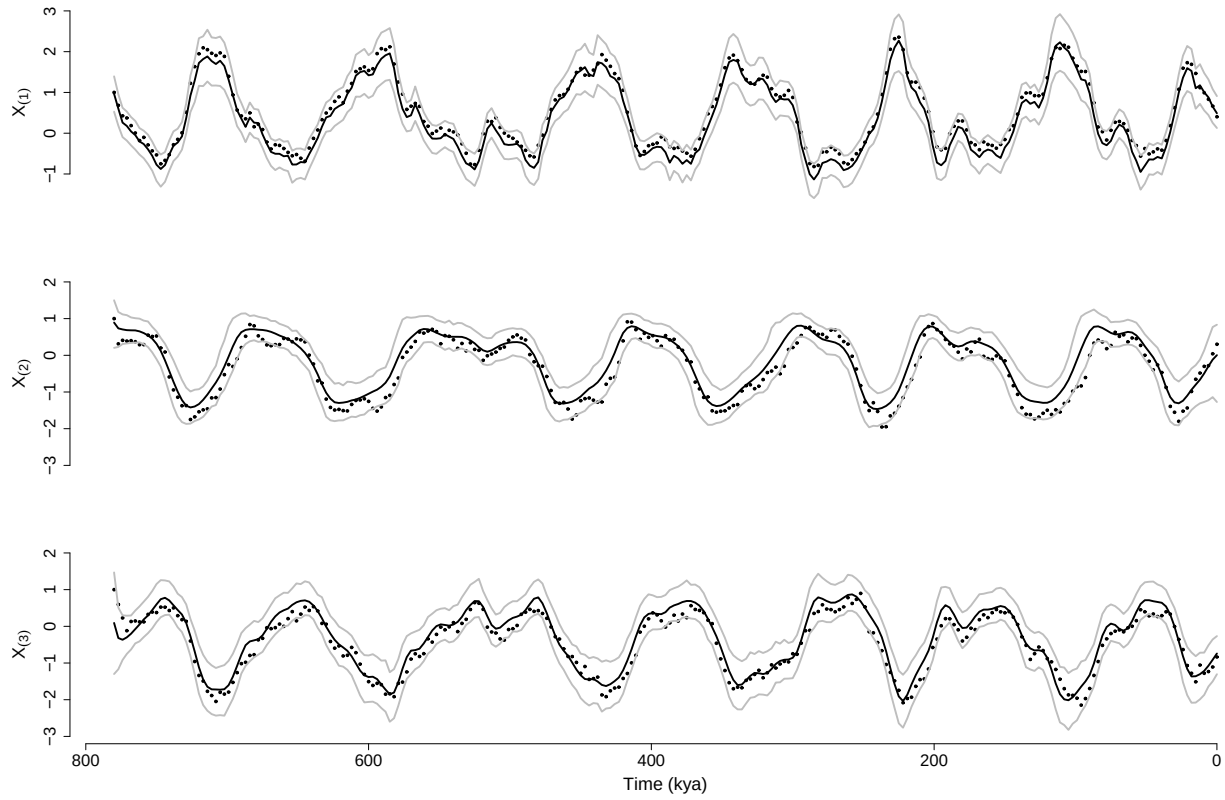

Fig. 4. Sequence of 95% highest density regions (HDRs) for the state of the forced SM91 model when fit to the SM91-f dataset. The black line shows the mean values of the marginal posterior distributions of the states, the grey lines show the 95% HDRs, and the points show the true generated values. Note that only $X_{(1)}$ is observable.

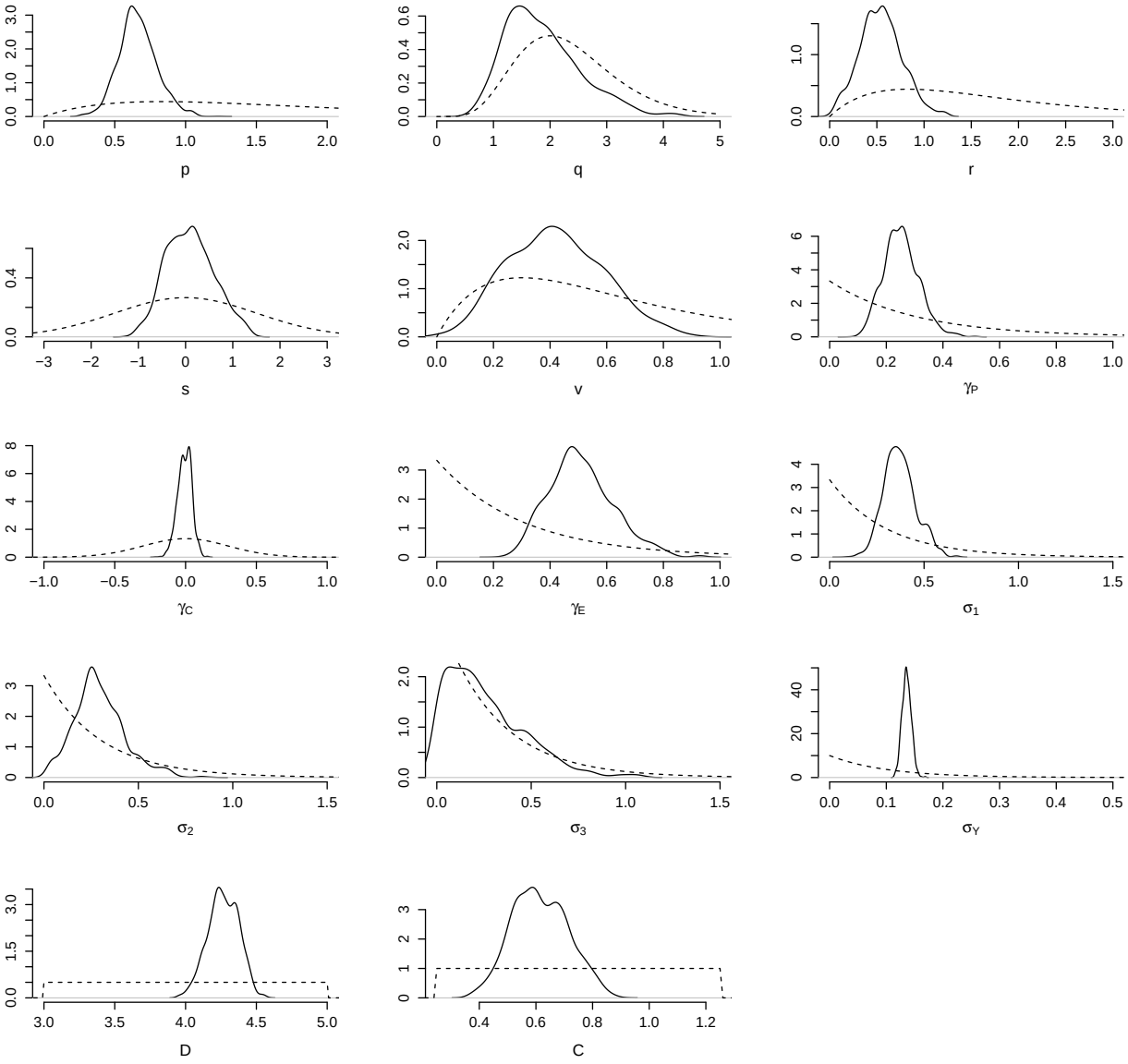

Fig. 5. Marginal posterior distributions for the fully forced SM91 model on ODP677-f. Dashed lines represent the prior distributions.

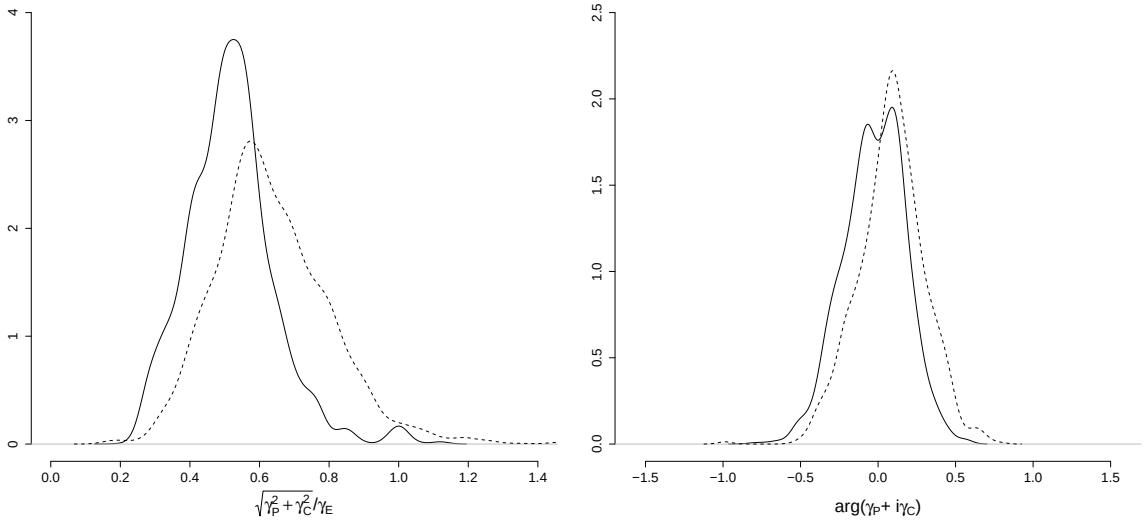

Fig. 6. Posterior density of the relative contribution between precession and obliquity in the astronomical forcing (left), and the phase of the precession (right) for the SM91 model (solid line), and T06 model (dashed line).

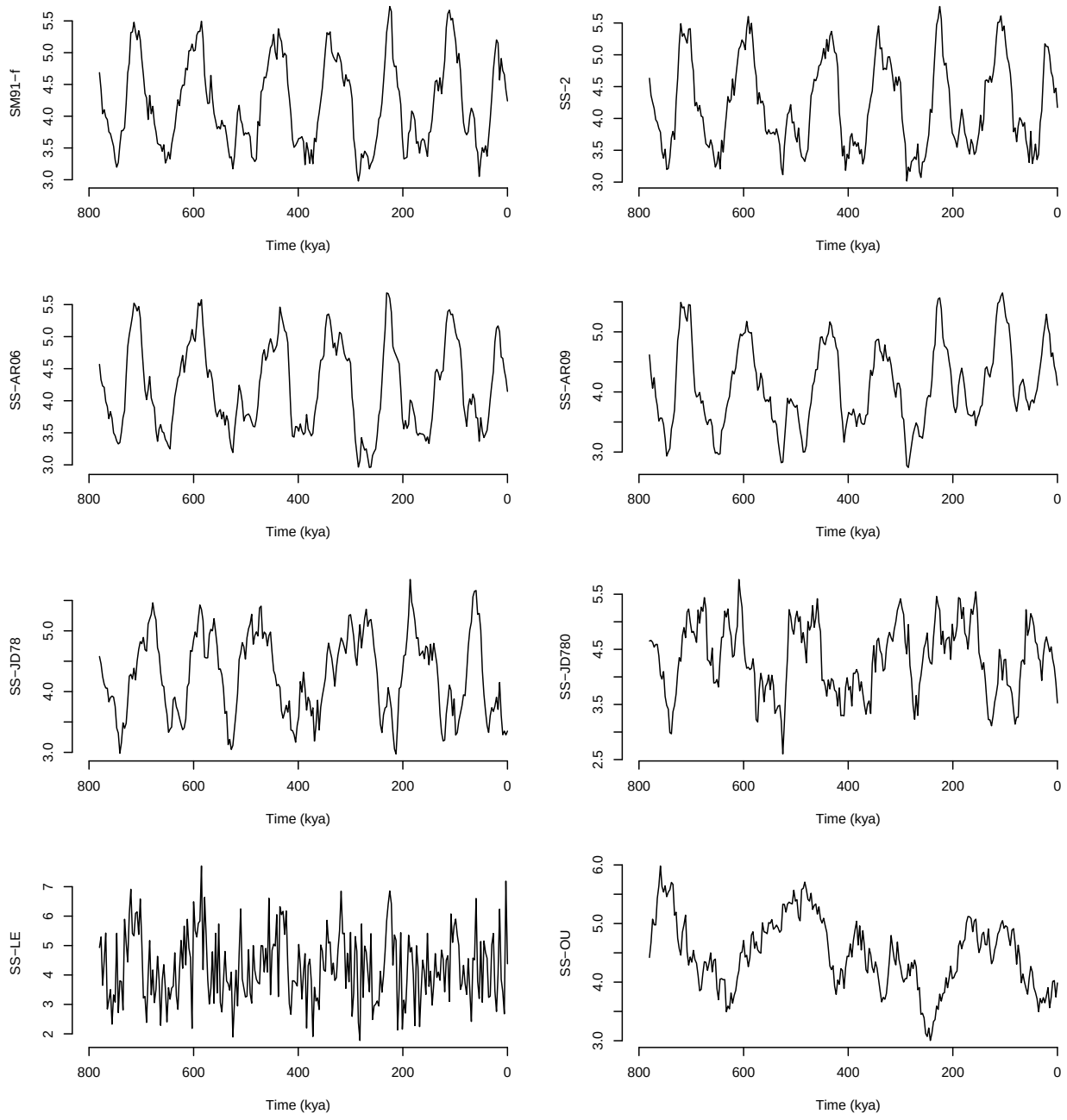

Fig. 7. Observed values in the forced simulation study datasets.